AL/HR-TR-1994-0013

ARMSTRONG LABORATORY

BUILDING A JOINT-SERVICE CLASSIFICATION RESEARCH ROADMAP: METHODOLOGICAL ISSUES IN SELECTION AND CLASSIFICATION

John P. Campbell, Editor
Teresa L. Russell

Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

February 1994

Interim Technical Report for Period January 1993 - December 1993

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS

NOTICES

This technical report is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

MALCOLM REE, Scientific Advisor
Aircrew Selection Research Branch

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Research Division

WILLARD BEAVERS, Lt Colonel, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE February 1994	3. REPORT TYPE AND DATES COVERED Interim January 1993 - December 1993

4. TITLE AND SUBTITLE

Building a Joint-Service Classification Research Roadmap:
Methodological Issues in Selection and Classification

5. FUNDING NUMBERS

C - F33615-91-C-0015
PE - 62205F
PR - 7719
TA - 24
WU - 03

6. AUTHOR(S)

John P. Campbell, Editor
Teresa L. Russell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory (AFMC)
Human Resources Directorate
Manpower and Personnel Research Division
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL/HR-TR-1994-0013

11. SUPPLEMENTARY NOTES

Armstrong Laboratory Technical Monitor: Malcolm J. Ree, (210) 536-3942.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Armstrong Laboratory and the Army Research Institute cosponsored a project to develop a Joint-Service classification research agenda, or Roadmap, for reducing redundancy of research across Services and improving inter-Service research planning. The Joint-Service Classification Research Roadmap is a research agenda designed to enhance the Services' selection and classification research programs. It is composed of numerous research questions that are organized into seven broad activities. Ordered roughly from highest to lowest priority, they are: Building a Joint-Service policy and forecasting data base, capturing criterion policy, modeling classification decisions, developing new job analysis methodologies, investigating fairness issues, conducting criterion measurement research, and conducting predictor-related research. The first two activities, "Building a Joint-Service policy and forecasting data base" and "Capturing criterion policy," will facilitate research planning. "Modeling classification decisions" and "Developing new job analysis methodologies" are activities wherein long-term research is needed. Classification is important because (a) changes in the ASVAB will result in revised composites, (b) recent innovations make classification research timely, and (c) downsizing makes classification more important. Job analysis research is needed to (a) facilitate innovations in predictor and criterion development and (b) facilitate management of selection and classification for future jobs. Fairness is important from a policy perspective. Criterion and predictor-related research are important, but the Services have researched them extensively. Extended research on experimental measures that have yielded promising results is recommended.

14. SUBJECT TERMS

Classification Measurement
Individual differences Roadmap

15. NUMBER OF PAGES

118

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

Table of Contents (Continued)

List of Tables

List of Figures

PREFACE

This technical paper documents research and development performed by the Human Resources Research Organization (HumRRO) for the Armstrong Laboratory, Human Resources Directorate, under Contract No. F33615-91-C-0015, JON 7719 2403. It is the fifth in a series of six reports to be delivered under this contract.

The Roadmap project products describe across-service military classification research issues. The key to the success of this effort has been the participation of experts from the Services. We thank the representatives of the Armstrong Laboratory, the Army Research Institute, the Navy Personnel Research and Development Center, the Center for Naval Analyses, the Defense Manpower Data Center, and the Office of the Assistant Secretary of Defense, who answered our questions. They were helpful and supportive. We are especially grateful to Dr. Malcolm Ree, who, as technical monitor of this contract, provided advice and technical support.

SUMMARY

The Armstrong Laboratory, the Army Research Institute for the Behavioral and Social Sciences,. the Navy Personnel Research and Development Center, and the Center for Naval Analyses are committed to enhancing the overall efficiency of the Services' selection and classification research. This means reducing redundancy of research across Services and improving inter-Service research planning, while ensuring that each Service's priority needs are served. With these goals in mind, the Armstrong Laboratory and the Army Research Institute co-sponsored a project to develop a Joint-Service classification research agenda, or Roadmap.

The Roadmap Project has six tasks. The first three tasks have been completed. Task 1 involved documenting the Services' current selection and classification practices and interviewing military selection and classification (S&C) experts to identify S&C research objectives (Russell, Knapp, & Campbell, 1992). Task 2 was a review of predictor measures (Russell, Reynolds, & Campbell, 1992), and Task 3 was a review and analysis of job analysis methods and procedures (Knapp, Russell, & Campbell, 1992). Task 4 reviewed criterion-related issues (Knapp & Campbell, 1992). The current report report discusses methodological issues in selection and classification.

The current report has six chapters. Chapter 1 provides an overview and describes the methods-related research objectives that emerged from Task 1. Chapter 2 discusses new methods for estimating gains due to classification and new procedures for making differental job assignments. Chapter 3 reviews methods for modeling the predictor and performance spaces. Chapter 4 focuses on synthetic validation, validity generalization, and mulitlevel regression prediction procedures. Chapter 5 describes methods of standards setting, and Chapter 6 discusses models and issues of fairness in testing.

BUILDING A JOINT-SERVICE CLASSIFICATION RESEARCH ROADMAP: METHODOLOGICAL ISSUES IN SELECTION AND CLASSIFICATıON

I. INTRODUCTION

Overview of the Roadmap Project

The Armstrong Laboratory, the Army Research Institute for the Behavioral and Social Sciences, the Navy Personnel Research and Development Center, and the Center for Naval Analyses are committed to enhancing the overall efficiency of the Services' selection and classification research. This means reducing redundancy of research efforts across Services and improving inter-Service research planning, while ensuring that each Service's priority needs are served. With these goals in mind, the Armstrong Laboratory and the Army Research Institute co-sponsored a project to develop a Joint-Service classification research agenda, or Roadmap. The roadmap project plan has six tasks:

Task 1.	Identify Classification Research Objectives,
Task 2.	Review Classification Tests and Make Recommendations,
Task 3.	Review Job Requirements and Make Recommendations,
Task 4.	Review Criteria and Make Criterion Development Recommendations,
Task 5.	Review and Recommend Statistical and Validation Methodologies,
Task 6.	Prepare Roadmap.

The first task, Identify Classification Research Objectives is reported in Russell, Knapp, and Campbell (1992). It involved interviewing selection and classification experts and decision-makers from each Service to determine research objectives. Tasks 2 through 5 are systematic reviews of specific predictor, job analytic, criterion, and methodological needs of each of the Services. The second, third, and fourth tasks are also complete and reported in Russell, Reynolds, and Campbell (1992) and Knapp, Russell, and Campbell (1992), and Knapp and Campbell (1992) respectively. This report fulfills the requirements of Task 5, Review Statistical and Validation Methodologies and Make Recommendations. The final task, Prepare Roadmap, will integrate the findings of Tasks 1 through 5 into a master research plan.

Research Objectives Related to Methodological Issues

Task 1 yielded a set of research objectives and information about military selection and classification experts' perceptions of the importance and urgency of those objectives. The objectives related to methodological issues are described below.

> **Investigate optimal strategies for incorporating predictor information into the assignment decision (e.g., alternatives for developing and using composites). (Objective #9)**

Incorporating predictor information into the assignment decision is the "bottom-line" in classification. Several kinds of research fall under this objective. For example, both validity generalization and synthetic validation suggest ways of incorporating predictor information into the assignment decision. Recent work by Johnson and Zeidner (1991) suggests new optimal classification methodologies might be promising. Also, a project currently being sponsored by the Air Force will examine the impact of adding new predictors to the ASVAB on classification efficiency. In this simulation study, American Institutes for Research (AIR) is manipulating several statistical and psychometric parameters of added variables (e.g., reliability of the criterion, the level of predictor-criterion correlation, amount of incremental validity) and using a linear programming technique to estimate the effects of various manipulations on classification efficiency. Several replications for cross-validation are included in the design to investigate the effect and magnitude of sampling errors.

Build an optimal assignment model that minimizes the impact of constraints on optimal assignment (e.g., "look-ahead" vs. strictly sequential processing to reduce impact of training slot availability). (Objective #12) and,

Increase the flexibility of assignment system (e.g., its responsiveness to supply and demand fluctuations). (Objective #13)

The next wave of technological advancements in assignment systems centers on: (a) improved methods of forecasting the characteristics of applicants, against whom the applicant is compared during sequential processing, (b) better ways of estimating and tracking the demand for personnel, and (c) user-friendly assignment software. Here, "look ahead" models forecast the supply of applicants and demand for recruits and identify the optimal combination of projected supply and job requirements. Look ahead processing can minimize the impact of constraints (e.g., training seat availability) and enhance the overall flexibility of the assignment system. Advanced algorithms that take maximum advantage of software/hardware technology will continue to improve our capability to classify recruits.

Investigate ways to maximize the influence of predicted performance in the assignment system (e.g., improve composite standard setting procedures; incorporate predicted performance into assignment algorithm). (Objective #14)

A primary goal of selection and classification is to maximize actual job performance of employees. However, because actual performance is unknown at the time of enlistment, mean predicted performance (MPP) is used as an estimate of actual job performance. Of course, the degree to which MPP is an accurate estimate of job performance depends on the sample sizes used to develop prediction equations and other sources of error in prediction. Moreover, MPP is not synonymous with actual job performance.

Several projects accomplished or underway address ways of maximizing the influence of predicted performance in the assignment system by linking enlistment standards to job performance. For example, Wise, Peterson, Hoffman, Campbell, & Arabian (1991) attempted a number of expert-judgment-based procedures for linking enlistment standards to job performance as a part of the Army's Synthetic Validation project. The most significant conclusions of this effort were that (a) Subject Matter Experts (SMEs) must fully understand the objectives and the consequences of the standard setting exercise to ensure the reliability and accuracy of judgments and (b) different methods of standard setting lead to different results. In short, further research is needed before Subject Matter Expert (SME) judgment based procedures will provide a viable way of linking enlistment standards to job performance.

The portion of the Job Performance Measurement (JPM) project that deals with the development of enlistment standards is called the Linkage project (Harris et al., 1991). The purpose of the Linkage project is to use JPM data (augmented with job characteristic information) to model the relationships among individual characteristics, job performance, and cost tradeoffs. The resulting model: (1) establishes the linkages between recruit entry characteristics, job performance, and costs, (2) permits better understanding and articulation of the tradeoffs between these dimensions, and (3) allows policy makers to set and revise entry standards based on efficient tradeoffs between performance and costs. The model aids the selection of the lowest cost mix of recruits (in terms of characteristics that are observable at the entry point) that can satisfy a specified performance goal and can be used to approximate the maximal performance and recruit mix obtained under certain budgetary constraints.

Evaluate alternative fairness models in terms of their effects on selection/ classification outcomes across subgroups (Objective #15). And,

Develop and evaluate extended models of fairness/equity issues by mapping out consequences of classification decisions at various stages in the selection and classification process. (Objective #16)

Adverse impact is "defined as a substantially different rate of selection in hiring, promotion, or other employment decision that works to the disadvantage of members of a race, sex, or ethnic group" (American Institutes for Research, 1992). Adverse impact is not, however, proof of unfairness. Cleary's (1968) model of fairness is currently accepted by both the Uniform Guidelines (1978) and the Society for Industrial and Organizational Psychology (SIOP, 1987). The Cleary model distinguishes between test bias and test fairness: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup" (Cleary, 1968, p. 115). In other words, a test is biased when prediction from a common regression equation results in either over- or under-prediction of subgroup performance. SIOP (1987) defines fairness as a social rather than a psychometric concept. Fairness is a function of how test scores are used for the job and the population at hand. For example, over-prediction of the performance of a protected group, when a common regression line is used, indicates statistical bias but is generally not considered a fairness problem.

Improve classification efficiency by improving strategies to generalize classification research findings across jobs and military populations. (Objective #19)

The problem the Services face in attempting to make predicted performance-based classification decisions is--how to construct prediction equations for literally hundreds of jobs that are constantly evolving and that range in population from very few to hundreds. There are really only two ways to form such equations without criterion-related validation data for every job--validity generalization or synthetic validation. Both synthetic validation and validity generalization strategies have shown some promise for generalizing classification research findings across jobs and military populations, but several questions remain unanswered. Both methods tend to underestimate validity, and synthetic validation procedures that capture differential validity need expansion (Wise et al., 1991).

The Linkage project, an expansion of the JPM project, also dealt with the development of a prediction equation in instances performance data are not available. Harris et al. (1991) used a methodology that best fits in the family of synthetic validation procedures. That is, a prediction equation was formed on the basis of the linkages between jobs, job descriptors (in this case, job characteristic variables) and, in turn, personal attributes (aptitudes).

Develop and evaluate alternative strategies and models for estimating the cost-effectiveness of an alternative classification system in terms of reduced training costs, reduced attrition, dollars, etc. (Objective #20)

With the draw-down of military manpower and equipment, resources have become even more of an issue. Diminished funding for personnel (testers, classifiers), places, and equipment places limitations on what is operationally feasible, and showing real cost savings is important for the implementation of new systems.

Previous attempts to show savings in terms of utility dollars have not proven highly successful in demonstrating to managers and executives the worth of new systems (e.g., CAT-ASVAB, Martin, 1992). The Defense Manpower Data Center's Concept-of-Operations Project (COP) which is currently underway will investigate strategies for estimating the cost-effectiveness of alternate selection and classification models.

Overview of the Current Report

This report has two principal goals: (1) to review critical methodological procedures and issues that are relevant for future selection and classification research and (2) to compare the issues addressed in the methodological literature to the issues incorporated in the relevant classification research objectives identified in Task One. These goals are addressed in a series of chapters organized around specific methodological issues.

There are perhaps four basic issues in selection and classification research. The first is how the goals of the decision system are represented. This is the criterion issue and it was addressed in a previous report. By default, or by design, the criterion that an organization uses to evaluate its selection and classification procedures are a reflection of its human resource management goals. However, any set of observed criterion measures is an imperfect reflection of the organization's goals (i.e., due to measurement error, deficiency, and contamination), and there are important methodological issues associated with determining the "fit" between the basic goals and the specific measures used to represent them.

A second issue is the identification of information that will best predict goal outcomes. That is, what predictors will forecast future criterion scores? The relevant predictor literature was reviewed in a previous report. However, there are a number of methodological issues associated with how the predictor information can be used most effectively and how the joint latent structure of predictors and criteria can be best represented.

The third issue is very critical for policy making and deals with how the degree of selection and classification accuracy/efficiency that actually exists in the population of interest can best be estimated. Given the complexities of the real world, this is a difficult statistical problem. It incorporates a number of sub-issues such as correction for estimation bias (e.g., restriction of range) and how to deal with assignment alternatives (e.g., jobs) for which there are no empirical estimates of regression parameters or validity coefficients. Given the complications, the overall issue is still the problem of estimation. That is, if the available predictor information is used to forecast the agreed upon criterion outcomes in the optimal (or sub-optimal) way, how accurate are the forecasts?

The fourth issue concerns the actual decision making procedures used to make classification assignments. How are scores on the predictors actually used to make job assignments given the constraints that everyone agrees must be satisfied? This is a mathematical, or decision modeling, problem, not a statistical issue. It is probably best represented by linear programming models. However, it incorporates at least two critical measurement sub-issues. The first concerns the determination of the maximum or minimum values (or "standards") on predictors or criteria that represent important constraints on the decision making system. For example, as determined by expert judgment, it may be stipulated that individuals in a particular job should not operate below a certain level of performance. This may stipulate in turn that, within some margin of prediction error, only individuals who are above a certain cut score on predicted performance can be assigned to the position. The second issue concerns the scaling of performance values (i.e., utility) so as to maximize the aggregate utility of job assignments rather than some other goal such as aggregate performance.

In addition to the above, there is another issue, which is primarily one of public policy but has critical methodological ingredients. It is the issue of fairness in selection and classification decision making, and how it should be modeled or represented.

This is a formidable list of methodological issues and by no means does the literature provide clear strategies for dealing with them. However, in the report we attempt to review what is currently known and to cast this information against a set of objectives for future methodological research and development.

II. NEW METHODS FOR ESTIMATING GAINS DUE TO CLASSIFICATION AND NEW PROCEDURES FOR MAKING DIFFERENTIAL JOB ASSIGNMENTS

Paul J. Sticha

Classification Objectives and Constraints

Efficient applicant assignment procedures should be controlled by the objectives and constraints of military selection and classification. Objectives define the functions to be maximized by the classification process. Constraints define the minimum standards that must be met by any acceptable classification solution. When a problem has both objectives and constraints, the constraints are of primary importance, in the sense that maximization of the objectives only considers candidate solutions that meet all constraints. On the other hand, additional capability beyond the minimum standard specified by a constraint adds no value to a candidate solution, while additional capability that helps to satisfy an objective better always improves the value of a solution.

When an optimal procedure, such as linear programming, is used to solve a classification problem, the objectives are represented as continuous functions to be maximized, while constraints are represented by inequalities among variables that must be satisfied. However, because of the duality between objectives and constraints in optimal assignment methods, it is in some sense arbitrary whether a particular factor is considered an objective or a constraint. For example, we could easily frame a classification problem as one of minimizing the total cost required to meet performance standards. In this case, minimizing cost would be the objective, while the performance standards would be the constraints on the classification process. Alternatively, the problem could be formulated as one of maximizing performance, subject to cost constraints. This approach reverses the role of objective and constraint from the first formulation. A third approach would maximize a function that combines cost and performance, such as a weighted average (the weight assigned to cost would be negative). This approach has no constraints, because both of the relevant variables are considered part of the objective.

It is possible for all three methods to arrive at the same solution, if the constraints, objective functions, and weights are set appropriately. However, in general, optimal solutions will satisfy the constraints exactly, or very nearly so, because objective and constraint variables are correlated. In the previous example, additional performance generally requires additional cost. Thus, if performance is considered to be a constraint in a classification problem, then the optimal classification will barely meet the performance standards, while minimizing cost. Thus, it is important to ensure that a classification that meets each constraint exactly is truly acceptable. If the constraints are set too low, then the optimization procedure may reach a solution that, in reality, is unacceptable. Alternatively, if the constraints are set too high, there will little room for the optimization of the objectives to occur.

Wise (in press) has identified eleven potential goals that could be addressed by selection and classification decisions, depending on the organization's priorities.

1. Fill available training seats with qualified applicants;

2. Maximize training success as measured by course grades;

3. Reduce attrition during the first tour of duty;

4. Maximize job proficiency measured by the percentage of job activities that the applicant could perform;

5. Maximize job performance, which incorporates measures of effort and discipline with proficiency;

6. Increase the qualified months of service, a term that which combines proficiency and attrition;

7. Improve total career performance, extending beyond the first tour;

8. Maximize performance utility, which incorporates the relative importance of different jobs;

9. Increase total job performance, which considers the entire distribution of performance in an job, including both its mean and variability;

10. Maximize unit performance/readiness, which considers groups of individuals with different jobs; and

11. Increase social benefit and avoid future social problems, which considers such aspects of the classification method as fairness to minorities.

This list illustrates the variety of goals that may be served by selection and classification processes. The individual goals are not mutually exclusive, but neither are they totally correspondent with each other, and no organization would be expected to try to optimize all of them at once. Some of these goals, such as maximizing training seat fill, are in close chronological proximity to the classification process, and can be easily measured. Others, such as total career performance, cover a period of time that may be many years removed from the classification process. In addition, some goals are at the individual level, while others are at the unit, job, or societal level. Finally, some are "vested" within others such as maximizing total performance utility which is really maximizing total performance, where levels of performance have been evaluated on a utility metric.

One important feature is that most of the goals on the previous list could be stated as either objectives or constraints. In addition, there are other constraints under which the classification system must operate. Probably the most obvious of these is cost. Other constraints are quotas for total accessions and for individual jobs, and minimum performance standards.

Different classification procedures focus on different objectives and constraints, and employ different methods to determine the optimal allocation of applicants to jobs. No existing method addresses all of the goals described above. The methods discussed in the remainder of this section are primarily concerned with the performance of people in their assigned jobs. Prediction of performance relative to objectives is a significant problem for classification methods. The discussion begins by describing methods to estimate the gains relative to objectives that can be obtained by classification procedures. The discussion continues by addressing methods for performing classification that attempt to maximize gains.

Estimating Classification Gains

Classification gains are linked directly to the goals of the classification process. That is, a better classification method allows an organization to meet it goals better. Both measuring the performance of existing classification systems and estimating the improvement from potential future systems present technical challenges. The extent of the challenge varies from objectives that cannot be measured in principle to objectives that can only be measured with considerable effort and cost. For example, it is impossible to measure directly how well someone would have performed in a job into which s/he was not placed (although the performance can be estimated from regression equations). Other objectives, such as job performance, can only be measured after a considerable amount of time after classification has taken place. A smaller challenge to measurement is presented by objectives related to performance during training or training seat fill rate.

Methods for estimating classification gains should be based on information about the classification procedure that is readily available, such as test validities and intercorrelations. Because of the complexity of the classification process, it may only be possible to estimate classification gains using analytical formulas when several simplifying assumptions are made. Under more complex and realistic assumptions, classification gains must be estimated using other methods, primarily those based on simulation.

Gains and Validity

Early measures of gains due to personnel procedures focused on selection rather than classification and were based on measures of test validity. The earliest measures of selection gains included Hull's (1928) Index of Forecasting Efficiency, defined in Equation (1).

$$E = 1 - \sqrt{1 - r_{xy}^2} \, , \tag{1}$$

where r_{xy} is the validity of test x in predicting performance on job y. Another early measure was the coefficient of determination, r_{xy}^2. Both of these indices express some pessimism regarding the utility of predictors with moderate validity. As illustrated by Zeidner and Johnson (1989), a predictor with a validity of 0.50 has a coefficient of determination of 0.25 and an Index of Forecasting Efficiency of only 0.13. As early as 1946, Brogden criticized these measures and demonstrated that r_{xy} is the proper measure of predictive efficiency.

The work on the value of selection and classification procedures starting with Brogden demonstrated that significant gains in classification utility are possible, even using predictors of moderate validity. Brogden (1955) provided the rationale for making classification decisions based on mean predicted performance (MPP), and for using a classification procedure based on full least square estimates (LSEs) of job performance from a battery of tests. He proved that the MPP will be equal to the mean actual performance for such a classification procedure, and that classification based on the full LSE composites produces a higher MPP than any other classification procedure. Several assumptions that limit the applicability of this result should be stated.

1. The regression equations predicting job performance for each job are determined from a single population of individuals. In practice, this assumption is infeasible because each individual has only one job. Consequently as Brogden states (1955, p. 249), "Regression equations applying to the same universe can be estimated through a series of validation studies with a separate study being necessary for each job."

2. There is an infinite number of individuals to be classified. Simulation research by Abbe (1968) suggests that the result is robust with respect to this assumption.

3. The relationships between the test scores and criterion performance are linear.

Brogden (1951, 1959) provided a method of estimating the MPP of a full LSE classification procedure, based on the number of jobs, the intercorrelation between job performance estimates, and the validity of the performance estimates. The development of this measure is based on the following assumptions.

1. There is a constant correlation (r) between each pair of performance estimates.

2. The prediction equations for each job have equal validity (v).

3. The population of people being assigned is infinite. This assumption is used to avoid consideration of job quotas.

Because the development of the method uses the results of Brogden (1955), those assumptions also apply.

From these assumptions, Brogden (1959) proved that the mean predicted performance (MPP), expressed as a standard score, is given by the following equation:

$$MPP = v\sqrt{1 - r}\ f(m), \qquad (2)$$

where *f(m)* is a function that gives the mean performance standard score as a function of the number of jobs (*m*) and the selection ratio. The function, *f(m)*, is based on the range of standard scores that would be expected in a sample from a normal distribution as a function of the sample size.

This result has several important implications regarding the determination of MPP. First, MPP is directly proportional to test validity. This result indicates that there can be much more value from tests of limited validity than was indicated by earlier estimates, such as the index of forecasting efficiency or coefficient of determination. Second, since MPP depends on $(1 - r)^{½}$, substantial classification utility can be obtained even when predictors are positively correlated. For example, Brogden (1951; adapted by Cascio, 1982) illustrated that using two predictors to assign individuals to one of two jobs can increase MPP substantially over the use of a single predictor (corresponding to a 0.17 increase in the standardized performance), even when the intercorrelation between the predictors is 0.8. Third, the results indicate that the MPP increases as the number of jobs (or job families) increases. The increase will be a negatively accelerated function of the number of jobs; for example, going from two to five jobs will double the increase in MPP, while going from two to thirteen jobs will triple the increase in MPP (Hunter & Schmidt, 1982).

The assumptions of equal predictor validities and intercorrelations are simplifications that allow for an easy, analytical determination of MPP. For more realistic cases in which validities and intercorrelations vary, MPP may be estimated using simulations. Procedures for conducting these simulations are discussed later.

Differential Validity[1]

The quality of predictors used for classification depends upon their ability to make different predictions of the performance on different jobs. That is, it should be possible, using the predictors, to predict with some accuracy that a person will perform better in one job than in another. Differential validity refers to the ability of a set of tests to predict the difference between criterion scores, such as performance on different jobs. Horst (1954) defined an index of differential validity (H_d) as the average variance in the difference scores between all pairs of criteria accounted for by a set of tests. It is not feasible to calculate H_d directly, because criterion scores are not available for more than one job for any individual. Consequently, Horst suggested substituting LSEs for the actual criterion scores.

[1] Differential validity has two meanings in the literature. Within the context of job classification, differential validity occurs when prediction systems make distinctions between jobs. In terms of fairness models (see Chapter VI), differential validity occurs when the observed validity coefficient for one group of people is significantly different from the observed validity for the second group.

Johnson and Zeidner (1990) have shown that when H_d is calculated based on LSEs and the assumptions used by Brogden (1959) hold, then H_d and MPP are related by the following function:

$$MPP = \sqrt{\frac{H_d}{m-1}}\, f(m). \qquad (3)$$

where m is the number of jobs, and $f(m)$ is defined as previously. Thus, when Brogden's assumptions hold, H_d and MPP are closely related concepts.

When H_d is calculated based on LSEs of criterion performance, then the index may be calculated from the matrix of covariances between the LSEs, denoted C. Specifically, the index is a function of the difference between the sum of the diagonal elements of C and the average off-diagonal element of C, as shown in the following equation.

$$H_d = \mathrm{tr}\boldsymbol{C} - \boldsymbol{1}'\boldsymbol{C}\boldsymbol{1}/m \qquad (4)$$

where tr $\boldsymbol{C}$ is the sum of the diagonal elements (or trace) of $\boldsymbol{C}$, $\boldsymbol{1}$ is a vector with each element equal to one, and m is the number of jobs.

Johnson and Zeidner (1989) have identified ten rules of thumb and measures that have been used to estimate gains from classification. They found that some of these measures have low accuracy. The measures, along with their accuracy as assessed by Johnson and Zeidner (1990) are shown in Table 1.

Recent research by Peterson, Owens-Kurtz, and Rosse (1991; Rosse & Peterson, 1991) used an index of "discriminant validity" that is the same as Rule 7, shown in Table 1. Specifically, the discriminant validity was defined as the "mean absolute validity minus [the] mean validity obtained by applying MOS equations developed for different MOS to a target MOS." The researchers used ASVAB scores to predict job performance using data collected for the Army's Project A. The results indicated very low discriminant validity, even though there were high levels of absolute validity.

Estimates of Gains from Simulations

The development of Brogden (1959) provides an analytic approach to predicting MPP when certain simplifying assumptions are met. When these assumptions are relaxed, analytical estimation of MPP is infeasible. Following methods used by Sorenson (1965), Johnson and

Table 2.1
Evaluation of Heuristic Rules for Evaluating Gains from Classification Methods

Rule Number	Description of Rule	Accuracy Rating
1	Composite (or test) intercorrelations	Low
2	Predicted performance intercorrelations	Medium
3	$R\ (1 - r)^{½}$	High
4	Predicted validity	Low
5	H_d	Medium to High
6	Comparison of diagonals of $\boldsymbol{V_a}$ with other row elements	Very Low
7	Comparison of diagonals of $\boldsymbol{V_a}$ with other column elements	Medium to High
8	Column variance of V	Medium
9	Dimensionality of either predictor or criterion space	Low
10	Dimensionality of joint predictor-criterion space	Medium to High

Note. Adapted from Johnson and Zeidner (1990). V denotes a matrix of test validities; V_a denotes a matrix of composite validities; R is the average multiple correlation between tests in a battery and each job; and r is the average intercorrelation between predicted performance measures.

Zeidner (1990, 1991), along with several of their colleagues (e.g., Statman, 1992), have applied simulation methods to examine the MPP of a variety of classification procedures. They have used these simulation procedures, which they call synthetic sampling, to test the predictions of a Differential Assignment Theory (DAT) of classification efficiency.

Basics of Differential Assignment Theory. There have been several summaries of Differential Assignment Theory, most notably those of Johnson and Zeidner (1990, 1991); Zeidner and Johnson (1989, 1991, in press); Johnson, Zeidner, and Scholarios (1990); and Statman (1992). Each of these descriptions presents somewhat different details of the theory. The following four points summarize the principles of Differential Assignment Theory.

1. **Importance of utility models.** Zeidner and Johnson (in press) stress the importance of using utility models to evaluate selection and classification methods. Utility models compare the gain in MPP associated with a particular classification process (relative to random classification) with the cost of using the process. Brogden's (1959) formula provides a method for calculating MPP if several assumptions are met. In most cases the MPP must be calculated through some kind of simulation procedure and cannot be calculated based solely on predictive validity.

2. **Use of full least square composites.** Full LSEs maximize the MPP for both selection and classification. Selection or classification methods that are not based on full LSEs optimize one function at the expense of the other. When full LSEs are infeasible because of the number of tests or jobs, then methods should be selected to maximize H_d (Horst, 1954). Constraints such as job quotas can be incorporated into the classification process through straightforward mathematical programming methods.

3. **The joint predictor-criterion space.** The joint predictor-criterion space is defined by the covariances of predicted job performance. According to differential assignment theory, optimal selection and classification methods are defined and evaluated in this space. A critical principle of Differential Assignment Theory is that there is a non-trivial degree of multidimensionality in the joint predictor-criterion space, corresponding to a general ability factor and job specific factors.

4. **Improving classification.** According to Differential Assignment Theory, substantial improvements in classification efficiency are possible through changes in the design of the selection and classification system. Improvements in classification efficiency can come from hierarchical classification effects or allocation effects. Hierarchical classification effects can improve classification efficiency even if classification is based on a single test by placing applicants with the highest predictor score in the jobs for which that score has the highest validity. Allocation effects require multiple predictors of job performance; benefits are derived by placing applicants in jobs that maximize their predicted performance. Increasing the number of tests, increasing the number of job families, and combining selection and classification into a one-step process, can all increase classification efficiency. Current computer technology has sufficient capability to implement any of these improvements.

Johnson, Zeidner, and Scholarios (1990) summarized the principles that form the basis of Differential Assignment Theory as follows:

> DAT provides a basis for generating a large number of principles applicable to the improvement of operational personnel systems. These principles are obtained as a result of focusing on the gains obtainable from a deliberate and methodologically correct attempt to capitalize on the differing requirements of jobs, using optimal selection and assignment algorithms in an appropriate context. This context includes: (1) appropriate test batteries; (2) best weighted selection

> and assignment variables; and (3) well structured job families. The psychometric principles of DAT are factual within the constraints of the assumptions necessary to derive them (p. 49).

<u>Estimating MPP with synthetic sampling</u>. Brogden's (1959) formula allows one to estimating the MPP of a classification procedure that is based on full LSE composites. However, this formula is not appropriate for estimating the MPP of classification policies that are not based on full LSEs. Furthermore, the formula will be inaccurate when the assumptions of equal validities and intercorrelations of the composites are not met. To estimate the MPP of a wider variety of classification procedures using more realistic assumptions, researchers have relied on a Monte Carlo approach termed synthetic sampling. Synthetic sampling allows one to estimate the MPP associated with any number of potential selection and classification policies. The basic approach is to evaluate classification methods based on random samples from a theoretical (multivariate normal) distribution representing the overall population test scores and job performance measures. Three classes of distributions are generated.

1. One sample is used to develop the prediction equations that form the basis of the assignment procedures. The assignment procedures may be based on LSEs, Aptitude Areas, or other combinations of the predictor variables. Several assignment procedures are developed from this sample, depending on the experimental design.

2. A second class of samples is used to apply the selection and classification procedures. These samples represent applicants who must be assigned to individual jobs, based on the procedures developed using the first sample. Usually several samples are made in this class. Each assignment procedure is used for each sample, producing a repeated measures design.

3. A third sample (or population distribution) provides the weights used to estimate the MPP for each of the assignment methods. In many cases, the weights are calculated directly from the parameters of the distribution that is used to generate the samples used for developing and applying the selection and classification methods. The population parameters, in turn, are inferred from empirical predictor intercorrelations and validity measures, corrected for restriction in range and criterion attenuation. The weights are applied to the scores of each simulated applicant to determine the performance in the job assigned by each classification procedure. In this way the MPP can be calculated for each of the samples (among the second class of samples) for each of the candidate classification procedures.

Analyses of synthetic sampling data compare the MPP for different assignment methods. The greatest MPP would occur if the population weights themselves were used to make assignments. Other assignment strategies produce lower MPP values for two reasons. First, the assignment weights are based on a sample from the population rather than from the actual population parameters. Second, all of the assignment strategies except those based on full LSEs are special cases of the optimal assignment strategy. That is, they reduce the number of factors considered or otherwise restrict the values for some of the weights of the composites used to predict performance. The variance of individual performance levels around the full LSE for the population enters into the analysis indirectly. If the variability is high, then the samples generated for development and application of the assignment methods will be dissimilar to the population values and to each other, thus producing lower MPP values. The synthetic sampling method assumes that the linear model is accurate. That is, there are no nonlinearities, and the distributions are all normal. Evidence reviewed by Hunter and Schmidt (1982) suggests that these assumptions are reasonable.

Results of the DAT tests. Table 2 summarizes key findings of studies investigating the implications of the DAT approach. All of the studies used the Project A data base. The ranges of improvements represent the range over several related experimental conditions. The data show the improvement in MPP in standard units, that is, as a proportion of the standard deviation of the MPP distribution. Each experimental condition was investigated with several synthetic samples (usually 20).

The researchers used the means and standard deviations of the MPP values, calculated over the 20 samples, to form the basis of statistical tests of the significance of improvements in MPP resulting from the experimental conditions, usually compared to current assignment methods (See Table 2 for specific methods compared). Standard errors are typically very small, and nearly all differences are significant.

The results were all consistent with the predictions of Differential Assignment Theory, although the magnitude of some of the results is modest. Full LSE composites lead to an increase in MPP of about 0.15σ when compared to current methods. Because of the size and complexity of the Army classification problem, this improvement is equivalent to a net present value of $260 million annually if, for purposes of illustration, the value for SDy is dollars is set equal to 40% of the average salary (Nord & Schmitz, 1989). Enforcing current Army quality distribution goals has little impact on MPP. Increasing the number of predictor tests, the number of job families, and the number of factors in composites, as well as decreasing the selection ratio all improve MPP substantially. The test selection method, job clustering method, and overall selection and classification strategy have much smaller effects.

These results represent potential improvements in classification efficiency, given a particular set of parameter values for things such as the number of jobs or job families, the dimensionality of the joint predictor criterion space, and the level of criterion intercorrelations for pairs of jobs. The actual improvement obtained by implementation of specific classification procedures will be less because of the damping influence of various constraints. For example, specific quotas such as the number of training seats available, will limit the extent to which individuals can be placed in jobs that maximize MPP, or assignment to the optimal jobs may leave training seats unfilled, thus increasing training cost. Alternately, if applicants have significant latitude in job choice they may not elect to take their best person-job match. The extent to which system constraints will reduce the expected performance gains below the

Table 2.2
Summary of Results from Model Sampling Experiments

Study	Methods Compared	Improvement in MPP (in standard scores)
Johnson, Zeidner, & Scholarios (1990); Scholarios (1992)	Test-selection method (H_d vs. Max-PSE)	-0.02 to 0.07
	Number of tests (10 vs. 5)	0.05 to 0.13
Nord & Schmitz (1989)	Full LSE composites vs. current methods	0.14 to 0.16
	Effect of Army quality goals on full LSEs	-0.01 to -0.02
Johnson, Zeidner, & Leaman (1991); Leaman (1992)	Number of job families (6 vs. 12)	0.09 to 0.13
	Job-clustering method (CE-based vs. operational)	0.03
Whetzel (1992)	Full LSE classification vs. g-based method	0.57 to 0.71
	One-stage select/classify vs. 2-stage	0.02 to 0.06
	Selection ratio (.50 vs. .75)	0.23 to 0.25
Statman (1992)	Number of factors in composites (8 vs. 1)	0.26 to 0.32
	Number of job families (10 vs. 4)	0.13

maximum possible or increase the cost required to obtain these gains is not known, and should be a future research topic.

The full maximization models are full LSEs that relate performance on a single job to all available predictors. All other models are reduced cases of this most general model. That is, they make restrictions on some of the weights in the full model. For example, they may restrict weights to be the same for jobs in a particular job family, or may restrict weights to have integer values, or may require that no more than two or three weights have non-zero value. Since the simpler models are all special cases of the most complex model, they cannot produce the MPP level that is obtained with the full LSEs, if models are calculated and evaluated in a single sample. The discrepancy may be large, or it may be small, but it will be evidenced to some degree. For example, the positive intercorrelations among predictors imply that MPP values will vary little with changes in the weights, but the expected value of MPP is always less than the MPP for the full LSEs. Statistical tests of these improvements are based on the sampling error introduced by the procedure for estimating gain in MPP based on multiple samples. These tests do not take into account the relationship between alternative models. Consequently, a statistically significant improvement in MPP will not necessarily be a practically meaningful improvement.

Utility Models of Classification Gains

Roach (1984) reviewed the use of decision-theoretic models for selection. He concluded that although there has been an increasing interest in utility models by personnel researchers, there has been little operational acceptance of decision-theoretic selection methods. More recent reviews by Zeidner (1987) and Zeidner & Johnson (1989) describe considerable advancements in decision theoretic models for selection and classification, but these reviews repeat the conclusion that there have been few applications of these models in operational settings.

Mean predicted performance is one measure of classification utility when the standard deviation of job performance does not depend on the job. In this case, the job performance standard scores will completely characterize the value of the importance. In other cases, job performance must be multiplied by some measure of the importance of the job in order to obtain a measure of the value of the performance. For example, Nord & Schmitz (1989) assume that utility (*Q*) is a weighted sum of performance across jobs. That is:

$$Q = \sum_{j=1}^{m} f_j(z_j, x_j), \qquad (5)$$

where z_j represents the job performance on job *j*, x_j represents other inputs, such as equipment and materiel, and f_j is a function that maps job inputs to a level of output, including consideration of the relative value of the job. Job performance is assumed to be a function of wages.

Nord and Schmitz (1989) developed a model to estimate the utility of selection and classification policies, including the Enlisted Personnel Allocation System (EPAS) and methods developed by Zeidner and Johnson (1989). They used two approaches to estimate a monetary value on the performance enhancements brought about by improved selection and classification. The approaches calculated the net present value (NPV) of the performance improvement and the opportunity cost, respectively.

The method for determining the NPV of performance applied and extended earlier approaches originally proposed by Brogden (1951), and extended by Hunter and Schmidt (1982) and others. NPV was defined by the following equation:

$$NPV = \sum_{i=1}^{N*} [\sum_{t=1}^{39} r_t(1 - A_{it})(P_i \times V_t - C_i^T)] - C_i^R, \qquad (6)$$

where N^* is the number of applicants that must be attracted to obtain the required number of qualified accessions, t indexes months, r_t is the discount factor used to determine the net present value (assumed to be 4%), A_{it} is the estimated probability that individual i will fail to complete t months of service, P_i is the expected performance in standard scores as predicted by full LSEs, V_t is the dollar value of one standard deviation of the performance distribution (assumed to be 40% of total salary), C_t^T is the monthly training cost, and C_i^R is the cost of recruiting an individual in the same ability range as i.

The results indicated a large improvement for both EPAS and the Zeidner and Johnson classification methods. NPVs were over \$50 million and \$260 million annually for EPAS and full LSE's, respectively. Most of the improvement was due to increased performance and reduced attrition. Because of the importance of predicted performance improvements, the results are very sensitive to the assumption that the standard deviation of performance, in monetary terms, is 40% of the total salary.

Because of the limitations of the assumption regarding the value of performance, a second analysis was made based on the concept of opportunity costs. The opportunity cost associated with a particular performance improvement is the expected cost required to obtain that improvement using the current system. Under the current system, performance improvements can only be achieved by increasing the number of high quality recruits. Increasing the number of high quality recruits will require additional recruiting cost; it may also affect the attrition rate. The opportunity cost can be estimated using the following formula.

$$OPPCOST_i = [(HQ_i \times ACH_i + (1 - HQ_i) \times ACL) \times (ACC + \Delta A_i)] - COST \tag{7}$$

where HQ_i is the required percentage of high quality accessions, ACH_i is the associated average cost of obtaining the high quality accessions, ACL is the cost of low quality recruits (assumed to be constant), ACC is the required number of accessions, ΔA_i is the change in attrition, and $COST$ is the recruiting cost under the current system.

The opportunity cost analysis showed an even greater value for the improvement brought about by EPAS and full LSEs, with assessed values of \$82 million and \$626 million, respectively. The large improvement from using full LSEs is based on the calculation that recruiting costs for high quality accessions would increase over 60% from \$8,371 to \$13,517. This estimate, combined with the estimate that 79% of the accessions would be required to be high quality to match the performance improvement produced by the use of full LSEs, lead to the high opportunity costs associated with these methods.

Making Differential Job Assignments

The previous section reviewed methods for estimating the degree of classification efficiency, or the degree to which a particular classification goal (e.g., MPP) can be increased as a result of a new classification procedure. The methods used to estimate the expected payoff in the population and the procedures actually used to make job assignments encompass two different set of issues. The estimation methods portray the maximum gain that can be achieved, given certain parameters, but the ability of the real-world decision-making procedures to realize the gain is another matter.

This section briefly describes the methods the services currently use to assign applicants to jobs, and the discusses research into the development of two future systems for making differential job assignments: the Army's Enlisted Personnel Allocation System (EPAS) and the Air Force's revised Processing and Classification of Enlistees (PACE) system. The overall goal of the new systems is to realize more of the potential maximum gained than is captured by the current systems.

Current Methods

All Services assign applicants to either an occupational area or a specific job at the MEPS. Although the process differs somewhat across the Services, generally a career counselor, or classifier, reviews the recruit's aptitude scores, medical history, and educational records. The counselor uses a computer system to obtain a list of current and future technical school vacancies and specialties, in order of Service priority, that match the applicant's records. Applicants and counselors discuss the job options, and the applicant makes the final decision about enlistment (Camara & Laurence, 1987).

Aptitude scores are an important component in each Service's assignment/classification system. Table 3 shows the names of the ASVAB composites used by each Service. Each Service has established minimum cut scores for each of its jobs or occupational areas on one or more of its composites to ensure a minimum level of aptitude for each job. Additionally, each Service uses aptitude scores to match people to jobs. However, the way in which this "match" is made and the type of information that goes into the "matching" process vary considerably by Service. The actual assignment of recruits to occupational areas or jobs is accomplished via computerized Person Job Match (PJM) algorithms. Each Service has its own algorithm, which reflects its current policies toward the relative priorities of filling jobs at any point in time. A brief overview of each algorithm is provided below.

Air Force allocation systems. The Air Force has two PJM systems. At the MEPS, the Procurement Management Information System (PROMIS) is used to make pre-enlistment assignments into either (a) specific jobs, Air Force Specialties (AFSs), through the Guaranteed Training Enlistment Program (GTEP), or (b) one of four occupational areas: Mechanical, Administrative, General, or Electronic (MAGE). Currently, about 30 to 40 percent of recruits are assigned into AFSs at the MEPS; 60 to 70 percent enter the Air Force with a guaranteed MAGE area. During Basic Military Training (BMT), recruits originally classified by PROMIS into one of the four MAGE areas are classified by the Processing and Classification of Enlistees (PACE) system into a specific AFS within the pre-assigned MAGE area.

Table 2.3
Current ASVAB Composites Used for Assignment by Service

Army	Air Force	Marine Corps	Navy	ASVAB Subtests
General Technical (GT)	General (G)	---	General Technical (GT)	AR + WK + PC
---	---	General Technical (GT)	---	AR + WK + PC + MC
Electronics (EL)	Electronics (E)	Electronics Repair (EL)	Electronics (EL)	GS + AR + MK + EI
Clerical (CL)	---	---	---	WK + PC + AR + MK
---	Administrative (A)	---	Clerical (CL)	NO + CS + WK + PC
---	---	Clerical (CL)	Business and Clerical (BC)	MK + CS + WK + PC
Motor Maintenance (MM)	---	---	---	NO + AS + MC + EI
---	---	---	Mechanical (ME)	AS + MC + WK + PC
---	---	Motor Maintenance (MM)	---	AR + AS + MC + EI
---	Mechanical (M)	---	---	GS + 2AS + MC
Combat (CO)	---	---	---	AR + CS + AS + MC
Field Artillery (FA)	---	---	---	AR + CS + MK + MC
Operators/Foods (OF)	---	---	---	NO + AS + MC + WK + PC
Surveillance/Communi-cations (SC)	---	---	---	AR + AS + MC + WK + PC
---	---	---	Basic Electricity/ Electronics (E)	GS + AR + 2MK
Skilled Technical (ST)	---	---	---	GS + MK + MC + WK + PC
---	---	---	Boilerman/Enginemen/ Machinist Mate (EG)	AS + MK
General Maintenance (GM)	---	---	---	GS + AS + MK + EI
---	---	---	Machinery Repairman (MR)	AR + AS + MC
---	---	---	Submarine (ST)	AR + MC + WK + PC
---	---	---	Communications Technician (CT)	AR + NO + CS + WK + PC
---	---	---	Hospitalman (HM)	GS + MK + WK + PC

Source: Bloxom (1992).

The Air Force assignment variables differ from those used by other Services in two ways. First, minimum physical strength requirements exist for many AFSs. Second, recruits indicate occupational preference by weighting (on a 0 to 9 scale) the M, A, G, and E areas. After all data are input to PROMIS, the program checks to ensure the applicant is eligible for the Air Force, identifies AFSs for which the applicant is eligible, and generates a relative payoff index (with a maximum of 1,000 points) that reflects the value of assigning the recruit to each AFS. PROMIS then compares the payoff index with the Air Force's current need to fill AFSs (based on training seat vacancies) and develops an ordered list of up to 16 AFSs. The first AFS is the "best choice" for both the individual and the Air Force (Pina, 1988). The specific functions that lead to the ordered list are summarized below.

Five components enter the PROMIS payoff algorithm to form the payoff index (with a maximum of 1,000 points): (a) variable fill versus aptitude/difficulty, 600 points, (b) predicted technical school success, 50 points, (c) (M, A, G, & E) area preference, 180 points, (d) minority/non-minority, 70 points, and (e) constant fill, 100 points (Pina, 1988). Variable fill is an index of the Air Force's needs at a particular point in time (i.e., number of personnel needed and the time remaining to fill the AFS). The aptitude/difficulty subcomponent matches individual aptitude to the level of aptitude required by the job (i.e., job difficulty). Variable fill and Aptitude/Difficulty interact such that aptitude/difficulty receives a larger allocation of the 600 points, if the Air Force's need for recruits is being met and vice versa. The technical school success component is based on regression equations for predicting technical school grades from AFQT, M, A, G, and E composites, and binary variables representing high school courses taken. The area preference component assigns points to M, A, G, and E areas in proportion to the applicant's preference. When PROMIS was originally developed the minority/nonminority component was designed to help meet the Air Force minority representation goals set for each AFS. Our most recent information is that the minority fill component still exists in the algorithm, but receives no points (L.T. Looper, personal communication, 14 April 1992). "Constant fill" is simply a constant of 100 points added to every AFS for which the applicant is eligible.

The current PACE is a simple, nonoptimal system that processes recruits in batch (i.e., non-sequential) mode (Pina, 1988; Pina, Emerson, & Leighton, 1988). It sorts recruits into available training seats on the basis of the recruit's (a) preference for the AFS, (b) ASVAB scores, and (c) gender.

Army allocation systems. The Army currently uses a computerized reservation, monitoring, and PJM system labelled REQUEST (Recruiting Quota System). A new assignment procedure, the Enlisted Personnel Allocation System (EPAS) was developed in a research effort known as Project B, but has not yet been implemented. The Army has no post-enlistment PJM system because specific jobs, Military Occupational Specialties (MOS), are guaranteed to all enlistees prior to enlistment.

The Army does not use job/occupational preference or physical strength variables for assignment, and aside from the gender exclusion policy prohibiting women from combat jobs, it has no minority fill component. REQUEST operates to achieve three goals: (a) to ensure a minimum level of aptitude in each MOS by applying minimum cut scores, (b) to match the distribution of aptitude within jobs to a desired distribution (i.e., to ensure a distribution of quality across jobs), and (c) to meet the Army's needs for filling MOS/training seats. Using functions related to these goals REQUEST computes an MOS Priority Index (MPI) that reflects the degree of match between the applicant and the MOS and uses the MPI to produce a list of MOS in order of Army priority. The functions involved in the MPI computation can be grouped into two broad categories: (a) MOS Status (MS) functions that define the Army's need to fill a particular MOS and (b) Applicant Qualification (AQ) functions that define the degree to which the applicant is matched to the MOS. The program first lists the five MOS that are highest in priority, and the classifier encourages the applicant to choose one of them. If the applicant is not interested in these jobs the next five high priority jobs are shown and so on until the applicant chooses a job. (Camara & Laurence, 1987; Schmitz, 1988).

__Marine Corps allocation systems__. Like the Air Force, the Marine Corps has two PJM systems for assignments. ARMS (Automated Reservation Management System) is used at the MEPS to assign applicants to either specific MOS or one of 35 occupational areas. Currently, only about two percent of recruits enter the Marine Corps with a guaranteed MOS. Nearly 85 percent are guaranteed an occupational area, and about 14 percent enter under an "open contract," with no occupational assignment. Most Marines are assigned to specific MOS after BMT; the Recruit Distribution Model (RDM) is used to make these post-enlistment assignments.

The Marine Corps uses its assignment system differently from the way in which the other Services use theirs. For most Services, occupational preference, if considered at all, is a piece of information in the algorithm with a known weight; the algorithm produces a list of options from which the applicant selects an occupational area or job. The Marine Corps relies more heavily on its counselors to assess job interests. Marine Corps applicants and classifiers talk about the applicant's interests. The classifier obtains a list of the applicant's preferences and calls the ARMS operator who, in turn, enters the applicant's data into ARMS. The ARMS operator checks to see whether the applicant can be assigned to his/her first preference. If not, the process is repeated until either a match is made or the applicant decides to enter the Marine Corps under an open contract. In short, occupational preference starts the assignment process. The ARMS algorithm ensures that applicants meet minimum qualifications for chosen MOS/occupational areas, fills available training seats according to Marine Corps priorities, and ensures that minority representation goals are met.

RDM is a batch-mode system used at Recruit Training Centers (RTCs) to assign recruits to job categories. RDM first fills jobs in accordance with the Marine Corps needs while meeting minority representation goals for jobs. After these two constraints are satisfied, the algorithm maximizes: (a) the average probability of success in training and (b) the number of recruits assigned to the highest prerequisite levels within each job category (Kroeker, 1989)

__Navy allocation systems__. The Navy's pre-enlistment assignment system, the Classification and Assignment within PRIDE (CLASP), works much like the Air Force's

PROMIS from which it was derived. At the MEPS, CLASP is used to make pre-enlistment assignments either into specific Navy jobs (or ratings) or into apprenticeship or general detail assignments. Currently, about two-thirds of Navy recruits are assigned to guaranteed training slots for specific Navy ratings. About one-third of the new recruits receive an apprenticeship or a General Detail assignment as Seaman, Airman, or Fireman. The Navy uses a post-enlistment system, Computer Assisted Assignment System II (COMPASSII), to assign recruits to ratings during BMT.

After data are input to CLASP, the program checks to ensure the applicant is eligible for the Navy, generates a payoff value reflecting the value of assigning the recruit to each rating, rank orders Navy ratings according to the payoff value, and eliminates ratings which have no openings or for which the applicant is otherwise not qualified. CLASP then presents the ordered list of Navy ratings for the applicant's consideration.

Six components enter the CLASP payoff algorithm: (a) predicted training success, (b) technical aptitude/job complexity, (c) Navy priority/ individual preference, (d) minority fill rate, (e) fraction fill rate, and (f) probability of attrition. School success is the predicted final grade based on ASVAB composite scores. Technical aptitude/job complexity is a numeric value for the expected relative utility of matching the level of individual aptitudes to the level of job complexity. Assignments that match on these two variables receive a higher value, and the value is proportionally higher if the match is for more complex jobs. Navy priority/individual preferences is an index of the relative value of assigning a recruit to ratings that vary in terms of the correspondence between the rating's Navy priority and the individual's preference. The minority fill rate component is designed to help the Navy meet minority representation goals for each rating. The fraction fill rate component evens the flow of allocations into ratings over the course of the recruiting month. That is, it gives utility points to ratings that have below average assignment rates. The attrition component is an estimate, based on demographic data, of the probability of retention during the initial service term and costs to the Navy for personnel loss (risk) for each rating (Kroeker, 1988, 1989; Kroeker & Folchi, 1984; Kroeker & Rafacz, 1983).

During the fifth week of recruit training, the Navy uses COMPASSII, in conjunction with a classification interview, to assign recruits to ratings. The interviewer recommends five occupational groups based on the recruit's ASVAB test scores, job experience, background, and preferences. After data are entered, COMPASSII conducts a series of optimization, each one constraining its predecessors. COMPASSII goals, in order, are: (a) to maximize the utilization of training seats, (b) minimize transportation costs, (c) match the interviewer's recommendations, and (d) maximize the probability of success in training schools (Hatch, Pierce, & Fisher, 1968; Kroeker, 1989).

Summary. There are similarities among the classification systems used by the Services. They all ensure adherence to minimum aptitude standards for each job, and all are designed to maximize the utilization of training school vacancies across jobs. Pina (1974) and Kroeker (1989) distinguish classification systems in terms of *how* they fill training seats. Systems that fill training seats (or vacancies) from available resources (within the constraint that each individual meets minimum job requirements) are "fill" oriented. "Fit" oriented systems match individual aptitudes and/or preferences to the jobs/available seats. The batch-mode, post-enlistment processing systems used by the Marine Corps (RMD) and the Navy (COMPASSII), for example, are driven primarily by fill policy (Kroeker, 1989). PROMIS and CLASP are examples of fit-oriented systems.

New Methods for Allocation

None of the current systems represent "true" classification in the sense that the entire set of job assignments is made such that the goal of classification (e.g., MPP) is maximized. All current systems seek to insure that one or more goals for each job are met even though the resulting assignments are sub optimal in terms of maximizing total gain. However, two new experimental systems have been developed which do incorporate a true classification component as part of the assignment algorithm. They are the EPAS system developed by the Army and the new PACE system developed by the Air Force.

EPAS. The Army's new, not yet operational system, EPAS, optimizes several functions simultaneously. They are designed to: (a) maximize expected job performance across MOS, (b) maximize expected service time, (c) provide job fill priority, and (d) maximize reenlistment potential. EPAS was designed to support Army guidance counselors and personnel planners (Konieczny, Brown, Hutton, & Stewart, 1990).

The following maximization problem provides a heuristic for understanding the view of the classification process taken by EPAS.

$$\begin{aligned} \text{maximize} \quad & Z = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} X_{ij} \\ \text{subject to:} \quad & \sum_{i=1}^{n} X_{ij} = 1 \\ & \sum_{j=1}^{n} X_{ij} = 1 \end{aligned} \qquad (8)$$

where the variables, i and j index the applicants and jobs, respectively. The matrix X_{ij} represents the assignment of people to jobs. If $X_{ij} = 1$, then applicant i is assigned to job j. The two constraints specify that each job is filled by a single applicant, and each applicant is assigned to a single job, respectively. The variable c_{ij} is a weight that represents the value or assigning applicant i to job j.

However, there are many factors that make the problem more complex than indicated in Equation (8), including sequential processing of applicants, applicant's choice of suboptimal assignments, multiple value criteria, complications caused by the Delayed Entry Program (DEP), and temporal changes in the characteristics of the applicant population. Consequently, the optimization approach taken by EPAS is considerably more complex than the simple formulation shown in Equation (8).

For example, in "pure" classification the optimal allocation of individuals to jobs requires full batch processing, but in actual applications, applicants are processed sequentially. EPAS attempts to deal with this complication by grouping applicants into "supply groups" defined by their level of scores on the selection/classification test battery and by other identifiers such as gender, educational level, etc.. For a given time frame the forecasted distribution of applicants over supply groups is defined and network or linear programming procedures are used to establish the priority of each supply group for assignment to each MOS. For any given period, the actual recommended job assignments are a function of the existing constraints and the forecast of training seat availability.

Consequently, the analyses performed by EPAS are based on the training requirements and the availability of applicants. EPAS retrieves the class schedule information from the Army Training Requirements and Resource System (ATRRS), and provides this information, along with the number of training seats to be filled over the year, to later processes. It then forecasts the number and types of people who will be available to the Army over the planning horizon (generally 12 months). The forecasts specify the distribution over applicant supply groups, based on recruiting missions, trends, bonuses, military compensation, number of recruiters assigned, youth population, unemployment, and civilian wages.

Based on the requirements and availability information, EPAS performs three kinds of analysis: (a) Policy analysis, (b) simulation analysis, and (c) operational analysis. The first two of these analyses are designed to aid personnel planners, while the third analysis primarily supports Army guidance counselors.

The policy analysis allocates supply group categories to MOS, including both direct enlistment and delayed entry. The allocation is based on a large-scale network optimization that sets a priority on MOS for each supply group. The analysis is used primarily for evaluating alternative recruiting policies, such as changing recruiting goals or delayed entry policies. The value used to determine the optimal allocation includes the expected job performance, the utility of this performance to the Army, and the length of time that the person is expected to stay in the job. Other goals include minimizing DEP costs, DEP losses, and training losses and recycles. Constraints include applicant availability, class size bounds, annual requirements, quality distribution goals, eligibility standards, DEP policies, gender restrictions, priority, and prerequisite courses.

The simulation analysis mode provides a more detailed planning capability than is possible with policy analysis mode. The simulation analysis produces detailed output describing the flow of applicants through the classification process. The simulation analysis may be based on the same network optimization that is used for policy analysis, or it may be based on a linear programming optimization. The linear programming model provides a more accurate representation of the separate requirements for recruit and initial skill training, and consequently produces a more accurate analysis. The linear programming model requires twenty times the computing time as the network formulations.

Operational analysis provides counselors with a list of the MOS that are best suited to each applicant. The primary differences between the operational analysis and the policy analysis are that the operational analysis allocates individual applicants to jobs, rather than supply groups, and performs sequential allocation of applicants. The module uses the lists of MOS provided by the policy analysis as the basis of its allocation procedure.

The ability of EPAS to "look ahead" derives from the interactions between the policy analysis over the planning horizon and the operational analysis. The policy analysis provides an optimal allocation over a 12-month period. This solution is one input to the sequential classification procedure used by the operational analysis. Individual assignments of MOS to an individual are scored according to how close they are to the optimal solution. Highly ranked MOS are those that are in the optimal solution. MOS that are lower ranked would increase the cost (reduce the utility) of the overall solution. The MOS are ranked inversely according to this cost.

The New PACE Payoff Algorithm. The Air Force has developed a new microcomputer-based PACE classification algorithm, but it also has not yet been implemented. The PACE algorithm includes the components of PROMIS, plus some additions. Supplementary PACE functions are designed to: (a) improve the fit between occupational preferences and assignment by improving occupational interest measurement, (b) take training costs into account, (c) minimize unproductive lag time (also called casual time) between BMT graduation and technical school entry, and (d) minimize first-term attrition. The ten components of the PACE algorithm are: (a) aptitude (M, A, G, and E composites), (b) job difficulty, (c) predicted technical school grade (based on ASVAB composite scores), (d) academic background (the percentage of desirable high school courses completed), (e) occupational interest (based on the Air Force Vocational Interest Career Examination, or VOICE), (f) restricted interest (the recruit's rankings of available jobs), (g) training cost, (h) the probability of retention during the first term of enlistment, (i) casual time (the number of days between BMT graduation and technical school entry), and (j) fill priority (the relative urgency of filling the AFS) (Pina, 1988; Pina et al., 1988).

The PACE payoff algorithm was developed using the "policy specifying" (Ward, 1977) approach that was also used to develop the payoff algorithm for PROMIS. The technique uses SME's, classification experts and policy makers, to define a post-enlistment classification policy for non-prior service airmen. The methods were designed to be similar to PROMIS and to avoid generating new data requirements. The algorithm is based on a person-job match (PJM) metric that combines the ten fundamental classification criteria organized into a hierarchical taxonomy. Six of the criteria address the effectiveness issues, that is, aptitude, interest, trainability, and so forth. The other four criteria are concerned with efficiency issues, such as cost, time, and fill priority. The first-level criteria are combined using the agreed upon combinatory functions to produce composite measures of effectiveness and efficiency. The effectiveness and efficiency measures are combined linearly to produce the individual's predicted score for each job; this score is used to make assignments that optimize the PJM. The relative weights given to effectiveness and efficiency in this combination are determined "by management at run time" (Pina, et al., 1988, p. 8). The assignment rules used to maximize the overall classification payoff (i.e., the PJM), are then computed using linear programming optimization methods. Since all information about both predicted goal outcomes (e.g., predicted training success) and constraints has been combined into one composite score, the solution for the optimal PJM becomes the familiar linear programming assignment problem.

PACE and EPAS differ in a number of respects. Perhaps the most distinctive is that EPAS attempts a simultaneous solution for the maximizing functions and constraint equations. PACE uses a much more compensatory model and combines almost all predictor and constraint information into one index before optimization takes place.

Cost-Performance Tradeoff Model. Research by McCloy, Harris, Barnes, Hogan, Smith, Clifton, and Sola (1992) developed a cost-performance tradeoff model that combines selection and classification functions. The goal of this model is to minimize the cost required to obtain a specified performance level. Thus, in contrast to other evaluation or allocation methods, predicted performance is considered a constraint in the model, and cost minimization is the objective. The model considers costs involved in recruiting, basic training, initial skill training, and compensation over the first term of enlistment. Performance estimates are weighted by the predicted survival probability, by month, over the first term. Thus, predicted attrition is incorporated in the estimation of both cost and performance. Separate submodels predict performance based on individual and job characteristics, calculate recruiting cost, estimate survival rates, and assess training and compensation costs.

The model uses quadratic programming methods to determine the selection and allocation strategy that minimizes the cost required to meet the required level of expected performance. One advantage of this method is that it does not require that performance be measured on a monetary scale. The optimization method can consider accession limits and quality distribution requirements. It does not consider other constraints, such as training seat fill requirements or casual time between basic training and initial skill training.

McCloy, et al. (1992) compared the prescriptions of the model to actual FY 1990 accessions for the Army and the Navy. They found that actual accessions were close to the values prescribed by the model. The estimated cost for the optimal policy was about 1% lower than that for the actual policy, leading to a predicted cost savings of $72 million for the Army and $31 million for the Navy. Considering quality goals had little impact on the cost of the optimal solution.

Comparison of Models. Both EPAS and PACE represent many of the goals of the classification process. However, the two methods represent these goals in different fashions. The PACE algorithm combines all goals into a single objective function which it then maximizes. EPAS treats predicted performance as the objective function, and other variables enter the model as constraints. (The cost-performance tradeoff model of McCloy, et al. (1992) does not consider the range of goals addressed by the other two allocation methods. However, it addresses cost specifically, in a way the other models do not. The many specific differences between the methods preclude a comparison of the pros and cons of the approaches. However, it would be possible to compare the methods using simulation studies.)

Research Issues

Effective classification must consider many objectives and constraints to assign recruits to the job where they can perform effectively, while maintaining the efficiency of the recruiting and training system. Differen' methods focus on different subsets of classification goals; no method addresses all goals. 'fwo major foci of recent classification research have been to develop classification methods that maximize MPP and to develop optimization methods to maximize classification objectives while satisfying constraints.

Recent research (e.g., Johnson & Zeidner) indicates that aggreeate job performance can be increased by incorporating classification methods that are concerned with maximizing MPP. However, research on these methods has concentrated primarily on *potential* improvements in classification efficiency, and has not yet addressed how well these methods operate under realistic constraints. It may be that the constraints on classification are so strong that all acceptable solutions produce very similar outcomes. In this case, there will be little to be gained from additional investments in classification technology. Research is needed to determine the extent to which constraints on making PJM assignments limit the gains that can be achieved. Such research would provide information that can be used to predict how much improvement is possible using assignment methods that try to capture as much classification efficiency as possible.

One constraint for which there is some knowledge is the Army's quality distribution goals. These goals ensure that there are sufficient high-quality applicants for future leadership positions and provide a hedge against uncertainty about future job requirements. Nord and Schmitz (1989) showed that meeting quality goals produced little loss in MPP. Consequently, these constraints are not particularly severe in the enlistment environment in which they were examined. However, this result should be interpreted in light of the relatively high proportion of high-quality recruits, who account for roughly two-thirds of all accessions. In an environment in which high-quality accessions were harder to obtain, the cost of meeting quality goals would be higher.

The potential benefits of using full LSEs as the basis of classification procedures, though large, may not be obtained in practice because job requirements change over time, changing the weights in the model from those that are used to make assignment decisions. Military downsizing is likely to change job requirements, producing fewer jobs with more varied duties.

The recommendation to base classification on full LSE composites runs counter to current procedures in one sense. Following this recommendation would have several implications that should be examined before the recommendation is implemented. The likelihood of negative weights in a composite would be unacceptable for selection, because the use of negative weights contradicts the instructions to the applicant to do his or her best on the test. Because classification is concerned with differential prediction, the issue of negative weights is much more complex. It can be argued that use of assignment composites that have negative weights can be unfair to applicants who are not assigned to their preferred job because they perform "too well" on a test that is weighted negatively in their preferred job. In fact, this situation can occur even if all weights are positive, as the following hypothetical example illustrates.

Suppose there are two jobs, J_1 and J_2, and two tests used for selection, t_1 and t_2. In this example, we assume that performance on J_1 is difficult to predict; consequently, both t_1 and t_2 have fairly low weights in the prediction equation. Performance on J_2, on the other hand, can be predicted very reliably from t_2 alone; consequently the weight for t_2 is high, while the weight for t_1 is near zero. Now consider an applicant who prefers J_1 to J_2. The predicted performance for J_1 increases with performance on both tests. However, if the applicant performs too well on t_2, then the predicted performance for J_2 will exceed the predicted performance for the preferred job. In this case, the applicant would be assigned to a less preferred job because of high test performance, even though all weights in the assignment functions are positive.

Problems of fairness of assignment algorithms should be an important concern in the development and evaluation of these algorithms. Negative weights are not required for problems with fairness to occur, although they increase the likelihood of such problems. It may be possible to restrict the values of weights to address concerns of fairness. The effectiveness of these restrictions in promoting fairness, as well as the extent to which they limit the benefits of the classification procedure should be examined.

Similarly, a full set of least squares estimates will include regression weights that are not statistically significant. That is, some of the weights in the prediction equation will not be significantly different from zero. The effects of including these weights in the classification procedure need to be examined in terms of their sensitivity to changes in job requirements, population abilities, and sampling error; as well as their impact on the fairness of the classification process.

What the above suggestions lead to is the need for a comprehensive sensitivity analysis of the effects of variation in critical features of the personnel system on (a) the ability of R&D to generate potential classification gain, and (b) the ability of various assignment procedures to capture the potential gain, given specific constraints. The available data now makes it possible to begin a more systematic modeling and evaluation of these issues.

III. METHODS FOR MODELING THE PREDICTOR AND CRITERION SPACES

Rodney A. McCloy

The goal of this chapter is to review methods now available for modeling the latent structure of predictor/criterion covariances. The methods will be discussed in terms of the specification of a latent structure that (1) explains the relationships among observed and latent variables and (2) may be tested empirically. A latent variable, or construct, is defined as "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Methods for establishing and testing a latent structure will be discussed. These methods often make great demands on subject matter experts and the data, but the potential payoff is substantial. Latent structures specify hypothesized relationships among psychological constructs and their operational definitions (i.e., the nomological network discussed by Cronbach and Meehl). Because of this close tie to theory development and testing, and the ability to explicitly account for measurement error in the observed measures of the constructs, modeling the latent structure of the predictor or criterion space is argued to be more beneficial to the understanding of psychological processes than mere empirical descriptions of observed variables.

The Latent Structure and Construct Validation

When modeling some portion of the predictor and/or criterion space, researchers examine the relationships among a set of observed variables. Sometimes they have specific hypotheses that we wish to test about expected relationships among the measures. At other times the endeavor is purely descriptive, the aim being to represent the relationships as accurately as possible. Whatever the goal, rarely does the interest lie in the observed measures, which are imperfect representations of the constructs they are designed to assess. Rather, the principal interest is in understanding the relationships among the constructs themselves. That is, the true objective of modeling is to describe and understand the latent structure of the predictor or criterion space.

Modeling observed variables by means of a latent structure is rooted in construct validation (Cronbach & Meehl, 1955). Through construct validation, researchers seek to determine (1) the degree to which the variance in an observed measure (e.g., a figural reasoning test, a work sample job performance test) is determined by the construct it was designed to assess (e.g., spatial ability, procedural knowledge/skill), and (2) the relationships among constructs.

Cronbach and Meehl (1955) discussed several methods for investigating construct validity, including studies of group differences, correlational and factor analysis, and studies of internal structure. Construction of a multitrait-multimethod matrix (D. T. Campbell & Fiske, 1959) has been perhaps the most commonly applied method. As James (1973) pointed out, the multitrait-multimethod methodology permits investigation of the relationships between the observed variables and the constructs they purport to measure (i.e., the epistemic definitions) but not of the relationships among constructs (i.e., the constitutive definitions). Specification of the connections between constructs requires other analyses (e.g., factor analyses). The relationships may be tested empirically, and different latent structures may be compared directly. The latent structure itself may be obtained either implicitly or explicitly. Once specified, the structure may be compared to sample data and its plausibility quantified. The following sections describe a method for generating a latent structure implicitly (policy capturing) and an empirical procedure for testing the fit of a model to sample data (structural equation modeling).

Implicit Generation of a Latent Structure--Policy Capturing Methods

The specifications for the linkages between the observed and latent variables constitute testable theoretical propositions. In turn, the results of the hypothesis testing lend support to or challenge the theory in question. Equating the latent structure to a set of theoretical propositions suggests that the researcher must explicitly formulate the structure to be tested, basing the variables it contains and the linkages among them on prior knowledge of the topic area. The latent structure of a group of predictors, or criterion measures, or their joint structure can be generated *a priori*, but this is not required. Indeed, in certain instances it may be better to develop the structure drawing on judgments of subject matter experts or members of the organization so as to reflect their policy in the network. Methods for obtaining policy information from relevant parties are known as policy capturing methods.

Policy capturing methods have been used in several military selection and classification projects. The following section describes some of the larger, more recent ones.

Validity Estimation--Army Project A. Policy capturing was used in the initial stages of Project A to identify the most promising predictor and criterion variables for the prediction and measurement of job performance (Wing, Peterson, & Hoffman, 1985). Following an exhaustive literature search, 53 promising predictor constructs were identified based on 12 evaluation criteria (e.g., reliability, group differences, test fairness). The constructs spanned a large portion of the predictor space and included cognitive (e.g., spatial, psychomotor, perceptual, verbal, and quantitative ability) and non-cognitive (e.g., temperament and interest) variables.

Criterion constructs were identified by reviewing descriptions of 111 jobs from 23 job clusters. Based on job activities and materials, 53 job-oriented performance constructs were formed (use maps in the field, control air traffic). Additional performance constructs were added that represented training performance (four constructs; e.g., effort/motivation in training) and general effectiveness (nine constructs; e.g., cooperation with supervisors). Finally, six other constructs were added--two required of all soldiers not considered in any of the other performance constructs (survive in the field, maintain physical fitness), and four constructs that are important to the Army and are outcomes of potentially many different behaviors (e.g., attrition, reenlistment). Thus, a total of 53 predictor constructs and 72 performance constructs were identified for study.

Three packets were formed, each containing descriptions of one-third of the predictors and all 72 criteria, as well as information on the concept of "true validity"--the correlation between a predictor and criterion devoid of the effects of range restriction, unreliability, or sampling error. The packets were then given to a group of 35 psychologists experienced in personnel selection research (e.g., researchers, professors). The judges then estimated the validity of each predictor for each criterion using a 1-9 rating scale (a rating of "1" representing a validity coefficient between .00-.10, "2" between .11-.20, and so on). Estimates were given for a predictor relative to all 72 criteria before the next predictor was considered.

A matrix of mean ratings of expected correlations for each predictor variable/criterion variable combination was constructed and analyzed by the method of principal components by columns (predictors) and by rows (criteria). These components represent higher-order latent variables for the predictor and criterion constructs, respectively. For example, there are many measures of dominance and finger dexterity.) The estimated validities describe the relationships among the constructs and hence represent connections in the latent structure.

Wing et al. (1985) demonstrated that judges knowledgeable of the variables in question can provide reliable and accurate judgments of the correlations among numerous predictor and criterion constructs, although the estimates showed a consistent tendency to underestimate values obtained from empirical research. The latent structure generated from the estimation procedure, once obtained, may certainly be modified. What should be stressed is that the method of obtaining estimates of validity coefficients from subject matter experts is useful should the goal of one's research be to examine as many variables as possible. Rather than imposing a certain structure on the data that might fail to consider important variables or relationships, an exploratory and descriptive approach is taken first that will suggest a starting point for theory building. This is certainly useful when exploring the predictor and/or criterion spaces, as were Wing et al. More specific and rigorous tests of the structure (both the variables it contains and the linkages among them) may be made at a later time.

Estimating Linkages To Form Equations--The Army Synthetic Validation Project. Using a validity estimation procedure, researchers in the Synthetic Validation Project (Wise, Peterson, Hoffman, Campbell, & Arabian, 1991) employed a validity estimation task to link Project A predictors to three different types of job component information (tasks, activities, or individual attributes). These estimates, in concert with information regarding the importance, difficulty, and frequency of various job tasks from the Army Task Questionnaire ("criticality" weights) and empirical estimates of predictor construct intercorrelations, were used to generate synthetic equations for predicting job-specific and Army-wide job performance. Various strategies for weighting (1) the predictors in the component equations, and (2) the component equations to form an overall equation, were investigated. The resulting equations may be generalized to other jobs for which no criterion data are available.

Two points should be made about the Synthetic Validation estimation tasks. First, the criticality weights contained the estimates of the importance of each job component for total performance. Second, results were much the same as from the Project A validity estimation study just described--the estimates can be made reliably and with reasonable accuracy.

Obtaining Weights for Constructing Composite Criteria--Project A. Suppose that one has factors or constructs representing dimensions of a higher order, multidimensional construct (e.g., Core Technical Proficiency and Maintaining Personal Discipline as dimensions of Army Job Performance). Although the individual factors are useful and meaningful in their own right, one still might wish to form a composite (e.g., Overall Job Performance) by weighting the various dimensions appropriately. If so, a policy capturing method could prove quite useful if one wished the composite to reflect the policy of members of the organization who evaluate individuals on the multidimensional construct. The policy of the organization's members would be reflected in the weights given to the components constituting the composite variable.

One procedure for estimating simultaneously the importance of various dimensions is conjoint scaling (e.g., Johnson, 1974; Green & Srinivasan, 1978). In conjoint scaling, judges must evaluate (e.g., rank order, rate) sets of stimuli that vary systematically with regard to the dimensions of interest. The weights given to the dimensions may be inferred from the evaluations of the stimuli. Stimuli may differ on all dimensions of interest at once (the full-profile conjoint analysis) or on two dimensions (the two-factor-at-a-time approach).

After examining several methods, the two-factor conjoint measurement approach was adopted by Sadacca, Campbell, White, and DiFazio (1989) to determine the relative importance weights to be assigned to the five performance constructs developed in Project A (Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Military Bearing; cf. J. P. Campbell, 1986). The five weighted constructs could then be summed to create a composite criterion of Overall Job Performance. The judges were NCO, company grade officers, and field grade officers, half coming from field units (FORSCOM and USAREUR) and half from proponent posts (TRADOC) from 20 Army jobs. Judges were presented with 10 sets of profiles on 15 hypothetical soldiers. All soldiers within a given set differed with respect to two of the five performance constructs. The judges were asked to rank order the 15 soldiers within each set in terms of overall job performance. Ratings were made within a military context of heightened tensions worldwide (i.e., a high risk of the breakout of hostilities). To the extent that soldiers who scored higher on construct A than construct B were ranked higher than soldiers who scored higher on B than A, A was taken to be the more important contributor to overall job performance than B. Sadacca et al. found that the pattern of weights given the five performance constructs varied significantly across the 20 jobs but did not differ with respect to the type of rater (e.g., NCO vs. field grade officer).

The conjoint scaling approach is another method by which the strength of the linkages between variables in a latent structure may be obtained. Conjoint scaling is quite empirical in that the weights applied to the dimensions of the composite are obtained through a scaling procedure based on the ratio of the dimension regression weights that are obtained when predicting a judge's rank ordering of the stimuli (Torgerson, 1958). If desired, however, one may eschew the empirical derivation of weights altogether. There are other means for deriving a single score from many dimensions that can be based entirely on judgment and policy.

Policy-Specifying--Development of the New PACE Classification Algorithm. The Air Force has two programs it uses to classify personnel. Classification prior to enlistment into specific jobs or into one of four Air Force Specialty (AFS) areas (Mechanical, Administrative, General, Electronic) is carried out using the Procurement Management information System (PROMIS). Classification from the four AFS areas to specific AFS is performed using the Processing and Classification of Enlistees (PACE) system. The PACE algorithm is "a mathematical model that uses information about the individual and the AFS to generate a payoff" (Pina, Emerson, Leighton, & Cummings, 1988, p. 5). The algorithm was developed using a procedure termed "policy-specifying" (Ward, 1977), "a decision-modeling technique by which variables identified as pertinent to a decision-making process can be combined to derive a single predicted payoff value" (Pina et al., p. 5). The predicted payoff value is the attribute used to classify individuals into jobs.

The first step was to form a panel of subject matter experts (classification experts, policy makers). The panel held weekly meetings during which it attempted to identify the most critical parameters for making assignment decisions. The process resulted in ten critical variables: aptitude, job difficulty, intellectual ability, academic background, objective interest, restricted interest, training cost, probability of first-term completion, casual time (the number of days between graduation from basic training and entry to a technical school), and fill priority. The structure of concepts generated from the panel discussions is presented in Figure 3.1.

The panel next selected measures of the ten variables. The measures were then combined into functions from the bottom up. For example, the measures of intellectual ability and academic background are combined to form a trainability score. Similarly, the ability score is a function of the trainability score and the aptitude vs. difficulty tradeoff score (which is a function of aptitude and job difficulty). The aggregation of measures continues until scores for effectiveness and efficiency are weighted and combined to form the score for the person job match. Classification decisions are then based upon this single index.

Unlike the component weights derived in the Sadacca et al. (1989) work which are recovered from ordinal (i.e. paired comparison judgments using a conjoint model, the PACE functions are analogous to direct magnitude estimation. For example, regarding the trainability function,

> The highest function payoff occurs when the scores for the two variables are each at their highest; the lowest payoff occurs whenthe scores are both at their lowest. The policy makers felt that intellectual ability is a more reliable indicator of trainability than is academic background; therefore, intellectual ability was given more weight in the function payoff (Pina et al., p. 9).

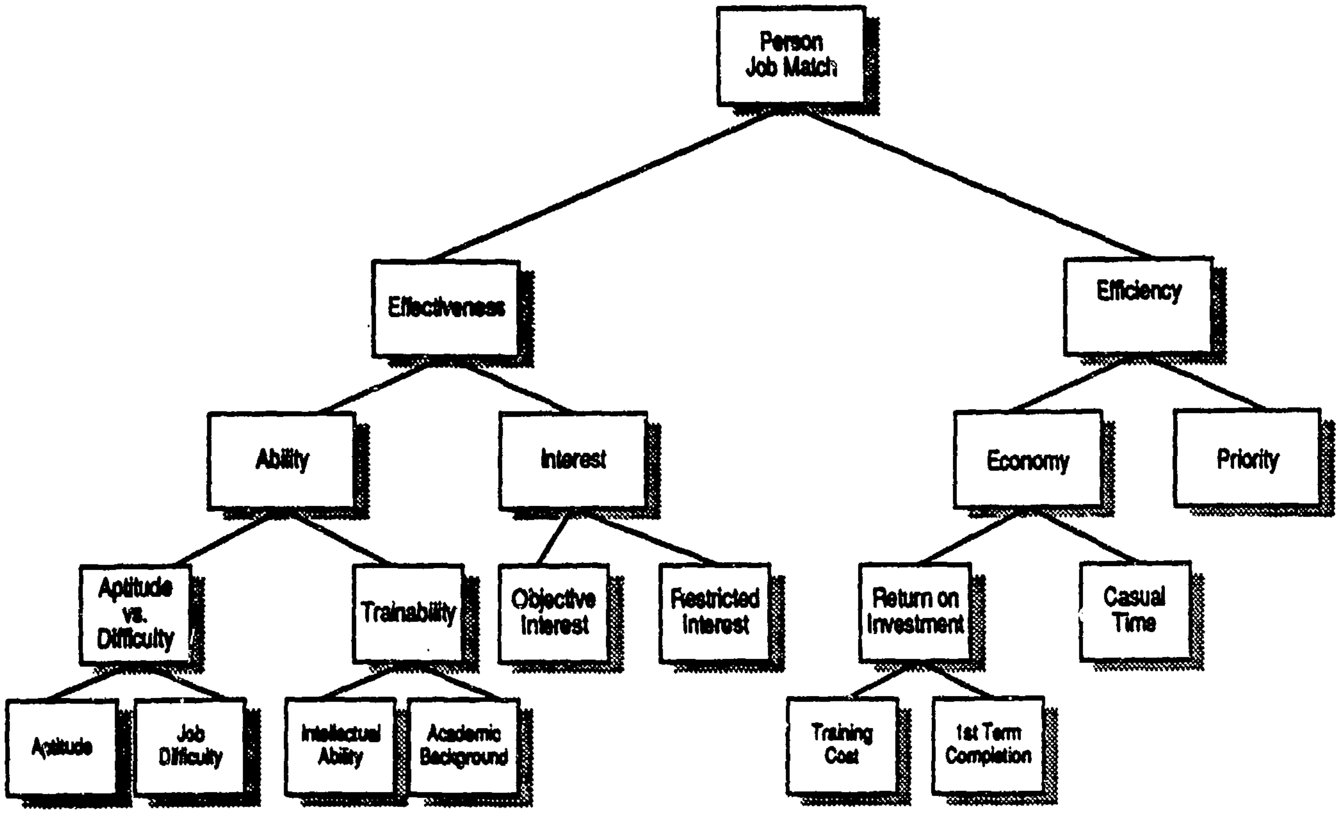

Figure 3.1
Conceptual taxonomy from PACE policy-specifying exercise.

The functions represent the policy of the subject matter experts, however, Pina et al. did not describe the procedure for how the specific importance values were obtained.

Summary. The methods described briefly above provide a means for specifying hypothesized linkages among variables in a nomological network. Policy capturing methods are of direct benefit when the decision rules are intended to directly reflect the policies of the organization. The weights given to the various variables under investigation may be derived by various procedures (e.g., conjoint scaling or policy specifying). They are also useful when conducting preliminary, exploratory, and/or descriptive analyses. The final result of any of these methods, however, should be the specifications for a hypothesized latent structure of the predictor and/or criterion space that may be tested for goodness of fit. The following section describes a powerful method for testing the theoretical propositions constituting a latent structure--structural equation modeling.

Structural Equation Modeling

Whether the latent structure is derived implicitly through one of the policy capturing procedures described above or explicitly defined *a priori*, the researcher must then move forward and test the hypothesized linkages between the observed and latent variables.

Causal Modeling. Relating observed variables to other observed variables can be accomplished through causal analysis (Wright, 1934; Asher, 1983). There have been only a few applications of causal modeling as a way of examining the latent structure of the predictor/performance space.

Hunter (1983, 1986) presented a path model specifying the causal relationships among measures of cognitive ability, job knowledge, job performance (operationalized as work sample performance), and supervisor ratings. The model stipulates the following relationships among the variables: (1) general cognitive ability directly affects job knowledge and job performance; (2) job knowledge directly affects job performance and supervisor ratings; and (3) job performance directly affects supervisor ratings. No direct path is designated between general cognitive ability and supervisor ratings.

Using 14 studies from both military and non-military settings having data for at least three of the four variables in the model, Hunter (1983, 1986) examined the fit of the path model to the average correlation matrix for each of the two settings. The model was found to fit the correlations quite well, and demonstrated "virtually perfect fit" (Hunter, 1983, p. 265) to the average correlation matrix resulting from the combination of the data across settings (from all 14 studies). Three major findings from his analyses are the following: (1) A substantial correlation was found between cognitive ability and job performance, "in part the result of the direct impact of ability differences on performance but . . . even more the result of the indirect causal impact due to the high correlation between ability and job knowledge and the high relevance of job knowledge to job performance" (Hunter, 1983, p. 265); (2) Supervisor ratings were more a measure of a ratee's job knowledge than of a ratee's actual job performance as manifested by the job sample measures; and (3) Job knowledge was a better measure of (was more correlated with) job performance (operationalized as work sample performance) than was a supervisor's rating of that performance.

Schmidt, Hunter, and Outerbridge (1986) expanded the Hunter performance model to include the effects of job experience. The coefficients for the model remained essentially the same. The effects of experience on supervisor ratings were of moderate size, most of the effect being indirect through job knowledge.

Finally, Borman, White, Pulakos, and Oppler (1991) applied the Hunter performance model to data from Project A and compared it to an expanded model that included multiple components of performance assessed via ratings, as well as non-cognitive predictor information (i.e., achievement orientation and dependability). Borman et al. found the expanded model accounted for over twice as much variance in supervisor ratings as Hunter's model. Cognitive ability, dependability, and job knowledge all demonstrated significant indirect effects.

Deriving Latent Variables. Although causal modeling yields the direct, indirect, and total effects of predictors on criteria (given certain assumptions that are often difficult to meet), the variables are usually observed measures, and researchers are typically not interested in the observed measures themselves. The Hunter (1983, 1986), Schmidt et al. (1986), and Borman et al. (1991) efforts just described all corrected the variables for attenuation.[1] Removing the measurement error from the variables is one way of approximating the modeling of latent variables, but this procedure makes the implicit assumption that removing measurement error yields a measure of the construct that is neither contaminated nor deficient--an assumption that might not be justified. For example, the supervisor ratings in these studies might contain systematic variance that is not related to the construct of performance (e.g., rating attractive persons higher than unattractive persons), resulting in a contaminated criterion variable (Brogden & Taylor, 1950). The presence of the contaminating variance can either inflate or reduce the relationship between the measure and another observed measure. Further, correcting the observed measure for unreliability does not remove this systematic variance. To the extent that such systematic variance exists in the observed measures, the goal of examining a causal model containing latent variables will not have been realized.

Rather than depending upon a single measure as an indicator of a construct, it may be possible to obtain several measures of the construct and to define the latent variable as the common variance among those measures. Factor analysis can provide insight into the latent variables accounting for correlations among a set of measures. Factor analysis alone, however, does not allow tests of the structural relations among the latent variables (i.e., the factors) apart from the estimation and testing of factor intercorrelations.

Integrating Path Models and Latent Variables--Structural Equation Modeling. The groundbreaking work of Karl Jöreskog in confirmatory factor analysis and the analysis of covariance structures (1966, 1967, 1969, 1970) led to a method combining confirmatory factor analysis with causal analysis. The method allows testing of an entire latent structure consisting of a factor structure for modeling the latent variables underlying sets of observed variables (indicators) and, if desired, the causal relations among the latent variables. Two models are defined and the model parameters are estimated simultaneously using one of several procedures (e.g., ordinary least squares, generalized least squares, maximum likelihood): a *measurement model* specifying the relationships among the observed variables (indicators) and the latent variables (factors), and a *structural model* specifying the relationships among the latent variables. The method, structural equation modeling, can be applied using the LISREL (LInear Structural RELations) software package (Jöreskog & Sörbom, 1989). Other packages are also available (e.g., Bentler, 1985).

[1] Borman et al. also provided models in which the criterion variable (i.e., supervisor ratings) was not corrected for unreliability.

Although the mathematics behind structural equation modeling is complex, the logic of the method is fairly straightforward. A theory about the latent structure specifies the hypothesized relationships among a set of observed and latent variables. This theoretical pattern of associations suggests a specific pattern of quantitative relationships (covariances). The general procedure for structural equation modeling is to obtain a set of sample data, calculate a covariance matrix among the observed measures in the sample, calculate the covariance matrix that the model (i.e., the latent structure) suggests, and compare the sample covariance matrix to the model covariance matrix. If the difference between the sample and model covariance matrices is small, then the sample data matrix is structurally similar to the matrix suggested by the model, and thus the model is deemed plausible for (i.e., the model fits) the data. If, on the other hand, the differences between the matrices is large, the model is said not to fit the data.

Specifying the Model. The relations among the variables in the hypothesized latent structure are specified using several parameter matrices (e.g., factor loadings, factor correlations). These model parameters may be free (estimated) or fixed (constrained to be a particular value, often zero). In addition, two or more parameters may be specified as free but constrained to be equal. The values of the parameters indicate the strength of the relationships among the variables. If a model's free parameters are a subset of a another model's free parameters, then the models are said to be nested. The fit of two or more nested models may be tested for statistical significance. Such tests give insights into the tenability of hypothesized linkages among variables. Generally, the more free parameters (i.e. to be estimated) a model contains, the more degrees of freedom it uses, and hence the better it fits the data. Models estimating few parameters are more restrictive and much easier to reject. However, if a very restrictive model is not rejected then the investigator has generated proportionally greater support for the model as an explanation of the latent structure.

Methods for Estimating Model Parameters. Parameters may be estimated using the method of instrumental variables, two-stage least squares, unweighted least squares, generally weighted least squares, diagonally weighted least squares, generalized least squares, and maximum likelihood (cf. Jöreskog & Sörbom, 1989). Each method requires certain properties of the data (e.g., maximum likelihood estimation procedures assume multivariate normality for the measures whereas generalized least squares procedures do not) and offers various advantages and disadvantages (e.g., unweighted least squares estimates are generated quickly and are consistent but not efficient; maximum likelihood estimates are consistent and efficient but often costly to obtain due to the iterative estimation procedure).

In addition, each estimation procedure is associated with a different fit function that is minimized by the program. For example, using unweighted least squares, the fit function to be minimized is

$$F = \frac{1}{2} tr[(S - \Sigma)^2]$$

where S is the sample covariance matrix, Σ is the model covariance matrix, and *tr* represents the trace (i.e., the sum of the diagonal elements) operator. Hence, the program minimizes the sum of the squared diagonal elements of the residual matrix, S - Σ. By comparison, the fit function to be minimized for maximum likelihood estimation procedures is

$$F = \ln|\Sigma| + tr(S\Sigma^{-1}) - \ln|S| - (p+q)$$

where ln is the natural logarithm, p is the number of endogenous variables in the model, and q is the number of exogenous variables in the model.

Assessing Model Fit. Many statistical indices are available for testing the fit of a hypothesized latent structure to sample data (i.e., the discrepancy between the model and sample covariance matrices). Perhaps the most commonly reported fit index is the chi-square statistic. The chi-square statistic is equal to (N-1) times the minimum value of the fit function, F. Although used as a measure of goodness-of-fit, the chi-square statistic is actually a "badness-of-fit" measure, since if it is "significant," the model is usually interpreted as implausible for the sample data. A more definitive statement is not warranted because of the substantial dependence of the chi-square value on sample size. If a sample is very large (e.g., N = 5000), then the power for the test is great and virtually any value will be significant, meaning virtually every model will not fit the data. In contrast, if the sample is small (e.g., N = 50), then the power to reject the null hypothesis is minimal and virtually any model will fit.

Another index of model fit given by LISREL is the goodness of fit index (GFI), "a measure of the relative amount of variances and covariances jointly accounted for by the model" (Jöreskog & Sörbom, 1981, p. I.41). This index typically ranges from zero to one, a value of one representing perfect fit. Negative values are nonetheless possible. This index may be used to compare the relative fit of models to different sets of data.

In addition to other measures of fit given by LISREL (e.g., the root mean square residual, which is the average of the fitted residuals and may be used to compare models fitted to the same data; the adjusted goodness of fit index; fitted and standardized residuals), numerous other fit statistics have been proposed for evaluating the fit of a model to sample data. For example, Browne and Cudeck (in press) recommended a point estimate of the root mean square error of approximation (RMSEA; cf. Steiger, 1990; Steiger & Lind, 1980), which they described as "a measure of the discrepancy per degree of freedom for the model." Perfect model fit is indicated by the lower bound value of zero. Unlike the chi-square and GFI, the RMSEA can increase as additional model parameters are estimated. Hence, it has the potential to reward more parsimonious models. Bollen (1986), Bentler (1990), and Browne and Cudeck (1993) have all suggested alternative measures of fit.

Structural Equation Modeling in Military Research. Structural equation modeling has been used in several high profile military research projects. Perhaps the primary advantage of this method is that it forces the researcher to specify and test explicitly a particular theoretical structure. It was used to model performance in Project A (J. P. Campbell, McHenry, & Wise, 1990) and the latent structure of an extensive test battery made up of traditional ability measures and a number of measures of cognitive processing developed as part of the Air Force's Learning Abilities Measurement Program (LAMP; Kyllonen & Christal, 1990).

Prior to the Project A work, there had been few attempts to theorize about what the latent structure of the performance space looked like. Two recent modeling efforts have provided additional insight into the performance space. Vance, Coovert, MacCallum, and Hedge (1989) proposed a latent structure of performance using performance criteria from the Air Force Job Performance Measurement effort. Four performance criteria (technical school grade; time to complete tasks on a work sample test; performance on a work sample test; and task ratings obtained from self, supervisor, and peers) are specified to be related to three classes of predictors (cognitive ability, experience, and supervisor support). The model was fitted to data obtained from three groups of tasks using LISREL. Vance et al. reported that "the model fitted marginally well in two of three cases" (p. 450). Substantial modifications were then made to the original model for each of the three categories of tasks, resulting in three models that are quite different from the originally hypothesized structure and each other. Note that some caution is advised since the modifications appear to have arisen primarily from empirical rather than theoretical considerations and the modified models were not cross-validated. However, for that part of the performance domain that seemed to be common across the army and Air Force performance measures, the latent structure described by the two modeling efforts seemed quite similar.

In contrast to the previous two studies, which attempted to model the substantive components of performance itself, McCloy (1990) proposed a latent structure for the direct determinants of performance and attempted to test it empirically. Using performance data from the Army's Project A, he hypothesized that the relevant (i.e., true) variance in a performance component (one or more job tasks that constitute a factor of job performance), as measured by different performance criteria (written tests of job knowledge, work sample performance tests, personnel file data, and peer and supervisor task ratings), is a function of the combined effect of three direct determinants that can be modeled as latent variables:

- Declarative Knowledge -- Knowledge of facts, rules, principles, and procedures. Specifically, declarative knowledge represents an the ability to state the facts, rules, principles, or procedures that are a prerequisite for successful task performance (Anderson, 1985; Kanfer & Ackerman, 1989).

- Procedural Knowledge/Skill -- The capability attained when declarative knowledge (knowing what to Do) has been successfully combined with knowing how, and being able, to do it (modified from Anderson, 1985, and Kanfer & Ackerman, 1989).

- Motivation -- As a direct determinant of performance, "motivation" is herein defined as a combined effect from three choice behaviors: (1) choice to expend effort; (2) choice of level of effort to expend; and (3) choice to persist in the expenditure of the chosen level of effort.

Hence,

$$PC = f\,(DK, PKS, M)$$

where PC is a particular job performance component, and DK, PKS, and M are the three performance determinants just defined.[2] Simply stated, the hypothesized performance function indicates that to perform a job task, a person must (1) know what the requisite job behaviors are, (2) be able to carry out the requisite behaviors, and (3) choose to carry out those behaviors for some period of time at some level of effort.

The model arose from a consideration of which determinants of individual performance differences could or could not be assessed by various kinds of criterion measures. For example, job knowledge tests are designed to assess what a person knows about how to perform a certain set of job tasks. As such, they appear to be direct functions of declarative knowledge. Because testing conditions are designed to assure motivation (as defined here) is maximal and constant for each individual, it most likely is not a critical determinant of individual differences in the test score. Further, PKS is not required to perform successfully on a job knowledge test (e.g., one need not be able to fly an airplane to do well on a written test of how to make a cross-wind landing). Similarly, work samples, in addition to assessing DK, are expressly designed to assess PKS. But given standardized conditions, motivation (as defined here) is again controlled. Finally, ratings and other measures of typical performance (Sackett, Zedeck, & Fogli, 1989) have the capacity to be influenced by all three determinants, because information regarding the volitional components of individual performance can be captured.

The hypothesized latent structure of the determinants of the various performance criteria suggests a simplex pattern of covariances among the criteria (Guttman, 1954; Jöreskog, 1970). Using a model for the analysis of covariance structures described by Bock and Bargmann (1966), McCloy (1990) conducted a confirmatory factor analysis using the Project A performance measures. The latent structure was confirmed across eight Army jobs. One major implication of the model is that if different determinants give rise to the observed variance in various performance measures, then correlations of those measures with another measure (e.g., a cognitive ability test) will also be different, <u>even if the performance measures assess exactly the</u>

[2] J. P. Campbell (1990) and J. P. Campbell, McCloy, Oppler, & Sager (1992) has significantly expanded the measurement model of determinants, marrying it to a taxonomy of job performance components to yield a theory of job performance.

same content. Thus, the substantive content alone of a performance measure is not sufficient to forecast its covariation with other measures.

Summary

Note that while a latent structure can be deemed plausible for a set of data; it can never be said to be the correct or "true" model. Indeed, several latent models, each postulating a somewhat different underlying structure, might fit a given set of sample data. Models may be compared on a relative basis, but they are admitted simplifications of complex processes and, strictly speaking, will always be incorrect. The goal of research is to accumulate evidence that puts a particular model (i.e. characterization of the latent structure) and its chief competitors to ever more stringent tests such that confidence grows that a particular model is a useful guide for research and for codifying research evidence.

After nearly a century of ignoring the latent structure of performance, the last few years have seen some beginning attempts to propose models of its basic nature. A theory of performance and a specification of its basic components seem necessary if research data on selection and classification are to be accumulated effectively and the value of a new selection/classification system for achieving alternative goals is to be evaluated meaningfully.

IV. DEVELOPING PREDICTION PROCEDURES AND EVALUATING PREDICTION ACCURACY WITHOUT EMPIRICAL DATA

Douglas H. Reynolds

As the United States entered the first World War, developers of the first group intelligence test sought a method of proving the value of their technique. After a period of definition and debate, E.L. Thorndike and his colleagues eventually reported on a procedure that depended upon the relationship between scores on the test and an external criterion (von Mayrhauser, 1992). Evidence of relationships between different tests and between tests and other criteria helped convince military officials of the usefulness of the intelligence measure. Since the validation of the Army Alpha, criterion-related validity evidence has been a critical component in the establishment of any applied testing program.

Although born of practical necessity, criterion-related validity is at the heart of the science of personnel psychology; a critical component of our understanding of work behavior is specifying the individual differences that influence it. However, we are only now beginning to witness progress toward the elaboration of a general framework relating psychological constructs and elements of performance.

This chapter presents recent developments related to the movement toward a general framework linking human characteristics and behavior on jobs and tasks. It has been observed that such a framework is a critical step in the scientific development of the field of personnel psychology (e.g., Guion, 1976; Peterson & Bownas, 1982), but it is still unclear what a general framework linking human characteristics and job performance would look like. Dunnette (1982) proposed a research agenda based on a hypothetical matrix of person-job characteristics, where the dimensions of the matrix are represented by a taxonomy of person performance components (abilities, personality traits, interests, etc.) and a taxonomy of job characteristics (e.g., tasks or work behaviors). The cells of the matrix would contain representations of the relationships between the person characteristics and performance components (e.g., correlations, variance percentages, judgments of overlap, or others measures of association). A detailed discussion regarding such a matrix was provided in an earlier report for this project (Knapp, Russell, & Campbell, 1992).

In the matrix described by Dunnette (1982), performance on job tasks is the criterion of interest. Recent efforts to define performance view the construct as multidimensional (e.g., Campbell, 1990), suggesting that an alternative view of the matrix would consider the various components of performance as another dimension. Job characteristics or other situational features (such as working conditions) may serve to define the context in which performance occurs and to moderate the relationship between individual characteristics and performance components.

The research presented in this chapter suggests that a general framework linking person characteristics, job characteristics and performance may be more complex than the matrix Dunnette (1982) proposed. For example, the relationships among person characteristics and performance components may be best described by validity distributions rather than individual estimates. The advances discussed in this chapter suggest that a general framework may need to be more elaborate than a two dimensional matrix; however, the notion that general rules can be established regarding person-job relationships has been extended by each of the efforts described.

Three areas of research and application are discussed here that provide support for the notion that criterion-related validity can be established without the collection of performance data in each situation in which a measure of individual difference is applied. These areas are synthetic validity, validity generalization, and the application of multilevel regression to performance prediction. First, we discuss the classic model of test validation, out of which each of the other topics has emerged.

Predictors, Criteria, and the Classic Validation Model

The traditional method of justifying test use for the prediction of performance is the demonstration of a relationship between performance on the test and performance on the activities to be predicted. This evidence has typically been developed through a "local validation study," in which a chosen predictor is correlated with a measure of later job performance. The predictive accuracy of the test may then be described as a correlation or as a function of that correlation.

Personnel researchers' dependence on the local validity study may have contributed to the pervasive belief in the situational specificity of test validities. "Situational specificity" refers to the notion that the criterion-related validity of a test is dependent upon the situation in which it is estimated. That is, a test found to be valid in one situation should not be assumed to be valid in other situations, no matter how similar. Local validation studies are often conducted with limited sample sizes and poor criteria, thus increasing the likelihood of obtaining results that would not generalize well across situations. Guion (1976) indicated that this situational focus in part impedes the development of generalizable rules that are essential for scientific progress in the field.

One roadblock to the development of generalizable test validity may be the tendency for researchers to focus on specific measurement methods and their relation to performance measures, rather than on the construct(s) the methods seek to measure. As argued in the previous chapter, focusing attention on the relationships between latent variables as opposed to the relationships between specific measures should improve the potential for generalizing validity. Furthermore, methods for generalizing test validity are only useful when adequate theories are available for describing the predictor space, the performance space, and job characteristics. Taxonomies of these areas are emerging, and a brief discussion concerning each area is provided below.

Predictor taxonomies. In an earlier report for this project (Russell, Reynolds, & Campbell, 1992), we described research on defining cognitive, psychomotor, personality, and interest constructs. Specifying the latent structure of the predictor domain should help uncover important and useful relationships between these variables and outcome constructs. For example, meta-analyses of the validity of personality measures have found relationships between specific personality constructs and performance (e.g., Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991), whereas in the past, reviews that grouped personality measures together into one domain were less likely to reveal useful relationships (e.g., Guion & Gottier, 1965; Schmitt, Gooding, Noe, & Kirsch, 1984).

Performance taxonomies. For many years, personnel researchers have lamented "the criterion problem" (e.g., unreliability and deficiency of criterion measures). At the same time, global judgments of performance have been the criteria in many validation studies. Only recently has concentrated effort been expended toward developing a better understanding of performance constructs. Campbell (1990) proposed an eight-factor taxonomy of performance that has been discussed elsewhere (Knapp & Campbell, 1992).

As the performance domain becomes better defined, more comprehensive and more relevant measures of performance are likely to be developed. For example, in the Army's Project A/Career Force research effort, a five-component model of performance was assessed with an array of performance measures. However, comprehensive performance measurement may be difficult to implement for many jobs. Thus, the capability to generalize relationships across jobs will be most valuable when multiple jobs in a group of similar jobs have extensive performance data.

The specification of the latent structure of performance is a critical step in the development of a general framework relating person characteristics and job performance. That is, as we become more exacting about which aspects of performance we expect to predict with a construct, we will be more likely to find stable and meaningful predictor-criterion relationships. In Project A, for example, personality measures were more strongly related to effort and leadership criteria than to other, non-volitional performance components. Had only a more global measure of performance been used, this pattern would not have been identified. It is important to recognize, however, the difficulties involved with the operational use of multiple criteria (cf. Cascio, 1992).

Job characteristics taxonomies. Taxonomies have been proposed for describing job characteristics such as variation in working conditions (cf. Peterson & Bownas, 1982). Several of these taxonomies were reviewed by Knapp et al. (1992). Job characteristics have traditionally been used to inform choices about the development and use of various predictors and criteria that are included in validation research (e.g., McCormick, Jeanneret, & Mecham, 1972). Alternatively, information about the characteristics of the situations in which a predictor is used may help to explain some variation in the criterion-related validities for that predictor. Situational information is also useful for generating estimates of predictor validity for jobs for which criterion data have not been collected.

Linkages between the latent variables. Traditionally, researchers have depended upon the local validation study to form empirical linkages between predictors and criteria. As mentioned above, however, reliance on local studies has led to unwanted consequences. Often local validity studies are conducted without enough data to reliably estimate the population parameter that describes the relation between the predictor and criterion. Thus, methods that estimate predictor-criterion relationships without data would be valuable because they offer an alternative to conducting studies that may produce inaccurate parameter estimates when acceptable amounts of data cannot be collected.

Several research strategies have been developed to estimate validities for jobs when traditional validation data are not available. The remainder of this chapter reviews three strategies: synthetic validation, validity generalization, and multilevel regression. It is important to keep in mind that these are simply methodologies for linking the various taxonomies; the more accurate the theory underlying each taxonomy, the better the resulting linkages from any of these procedures should be.

Synthetic Validity

Lawshe (1952) defined the basic logic behind synthetic validation when he proposed the technique as a method for developing a valid battery of tests in situations where conducting a criterion-related validity study would be impractical (e.g., in a small business). The rationale for the procedure is straightforward: determine the validity of a set of individual characteristics for predicting performance on a number of job components, then assemble a test battery for any given job based on the components of the job. Implicit in this rationale is the notion that different job components are best predicted by different human attributes.

In their review of the synthetic validation literature, Crafts, Szenas, Chia, and Pulakos (1988) described the process as consisting four basic steps. First, a taxonomy of job content is used to define a set of job components that can describe each of the jobs of interest. It is critical that all of the jobs under consideration be described in terms of the same components. Second, the importance of a number of human attributes for performing each of the job components is determined. Third, job components are identified that are important for each of the jobs. Fourth, information about the importance of each component and the relationships between predictors and components are combined to form a job-specific prediction equation. A number of different approaches have been taken for executing the synthetic validation process. A brief review of these approaches is provided below.

Approaches to Synthetic Validation

Although synthetic validation has been discussed for some time, it has been neither extensively researched nor consistently applied. Thus a number of different approaches have been developed. The approaches often vary in terms of how the estimates of validity are made between individual attributes and job components. A few of these approaches will be considered;

other approaches are detailed in several literature reviews that are available on the topic (e.g., Crafts et al., 1988; Hollenbeck & Whitemer, 1988; Mossholder & Arvey, 1984).

The J-Coefficient. The J-coefficient is a numerical index, derived through synthetic validation procedures, that expresses a test-job performance relationship in much the same way as a validity coefficient does (Primoff, 1955). The J-coefficient is computed by identifying a set of "job elements" (knowledge, skills, abilities, etc.) that are common both to a test and to job performance. The J-coefficient is then expressed as a function of the degree to which the elements are (1) important for performance and (2) measured by the test.

Element-job relationships can be estimated subjectively via expert judgment, or empirically by correlating current employees' performance on the elements to their overall job performance (Mossholder & Arvey, 1984). Element-test relationships are typically estimated by having test experts judge the relevance of each test item for each element, or these relationships can also be estimated empirically by correlating element proficiency ratings with test scores. Although it is possible to derive each of the necessary relationships empirically, the amount of data needed to produce stable estimates often precludes the method, so judgments are most often used.

Research comparing J-coefficients to traditional validity coefficients indicates that the two estimates are often quite similar (Dickinson & Wijting, 1976). However, missing critical job elements may lead the J-coefficient to underestimate validity (Mossholder & Arvey, 1984), and variation in element intercorrelations or test-element correlations across jobs will reduce the generalizability of the equations across jobs (Trattner, 1982). Nevertheless, the J-coefficient represents one way of developing an estimate of test validity without collecting criterion data, but it has been suggested that the J-coefficient may be more akin to content validity than to synthetic validity (Guion, 1976).

PAQ attribute profiles. The Position Analysis Questionnaire (PAQ; McCormick et al., 1972) is a worker-oriented job analysis questionnaire consisting of almost 200 general performance behaviors. McCormick et al. (1972) sought to establish the ability requirements for each dimension of performance behaviors by having a sample of psychologists rate the importance of each of 68 human attributes for performing each behavioral component.

The PAQ provides a method for completing several of the major steps for establishing synthetic validity. First, the PAQ provides a behavioral taxonomy that can serve as a basis for job analysis. Second, the established attribute ratings can be used to determine the ability requirements for the behavioral components. Third, jobs can be analyzed using the PAQ to determine the relevant behavioral components of the jobs and thus their likely ability requirements. However, the procedure does not provide a method for combining information into job-specific validity estimates (although McCormick et al. did predict validities for the General Aptitude Test Battery (GATB) with the ability requirement judgments). Thus, the attribute profiles may be more useful for informing a choice of tests rather than for estimating the validity of those tests.

An empirical process for establishing synthetic validity has also been used with the PAQ, however the procedure has been criticized on several grounds (cf. Mossholder & Arvey, 1984; Crafts et al., 1988). A description of the procedure appears in an earlier report (Knapp et al., 1992).

<u>Army SYNVAL project</u>. One of the primary goals of the Army's SYNVAL project was to determine what information should be used to select and classify people into an MOS when empirical validation data are not available (cf. Wise, Peterson, Hoffman, Campbell, & Arabian, 1991). Specifically, the project investigated the feasibility of using synthetic validation procedures for extending the results of the Army's Project A to Military Occupational Specialties (MOS) that were not included in that research effort. Again, a description of this research also appears elsewhere (Knapp et al., 1992).

The SYNVAL project investigated a number of issues dealing with synthetic validity procedures. The project compared different job component taxonomies (tasks, behaviors, attributes), synthetic vs. empirical prediction of specific performance dimensions (Core Technical, General Soldiering, and Overall Performance importance), different types of judges, and different weighting strategies for developing prediction equations.

SYNVAL research found little difference among the three component taxonomies in terms of their usefulness for analyzing jobs for synthetic validation purposes. Tasks, behaviors, and attributes could each be rated reliably in terms of their relative importance for measuring or predicting performance in an MOS, given a reasonable number of judges (approximately 15). Task components were chosen as the preferred unit of analysis for two reasons. First, using task components to describe jobs yielded greater discriminability across MOS than the other models. This was especially true when the each task component was rated in terms of importance for the Core Technical performance factor in each MOS. Second, rating the importance of tasks for an MOS tended to be more acceptable to judges (Army NCOs and Officers) than the other methods.

Importance ratings of task components for different performance dimensions indicated that there was a high degree of correspondence among ratings of a specific task's importance for the Core Technical, General Soldiering, and Overall Performance scales. Task importance ratings for General Soldiering showed the least discriminability between MOS, and Core Technical importance ratings showed the most discriminability. This was the expected pattern because General Soldiering activities are common across MOS, and Core Technical activities were defined to be MOS-specific.

The research also found that raters of varying supervisory levels and organizational backgrounds (i.e., training vs. operation) provided similar judgments of task importance. Attribute validity judgments for each task component were also made reliably by raters with at least some experience with psychological measurement. Different methods of weighting judgments of validity also did not produce large differences in the resulting equations, but a weighting strategy that applied a weight of zero to validities that were estimated to be low (e.g., below .30) slightly improved the discriminant validity of the resulting equations.

The SYNVAL project also compared synthetically derived validity equations with empirically derived equations developed for jobs where criterion data were available. The synthetic equations achieved levels of absolute validity (corrected for range restriction and criterion reliability) that were about 96 percent as high as the empirical equations. However, the empirical equations evidenced somewhat greater differential validity than the synthetic equations. Differential validity in this context refers to the difference between (1) the validity obtained for an MOS using its own equation, and (2) the validities obtained using equations calculated for the other MOS. Differential validity was found to be slightly higher when Core Technical performance criteria were used, compared to Overall Performance criteria.

Issues and Conclusions Regarding Synthetic Validation

As the SYNVAL research noted, two basic criteria for judging the usefulness of synthetic validation procedures are overall criterion-related validity and differential validity. Regarding overall validity, synthetic equations should be able to approximate the validity estimates derived empirically. Research on the J-Coefficient (Dickinson & Wijting, 1976) and the SYNVAL project (Wise et al., 1991) indicates that this condition can be achieved.

Differential validity evidence from SYNVAL indicated little difference between equations that were developed for one job and then applied to another. Differential validity is an important criterion for synthetic validity because the procedure is based, in part, on the assumption that different human characteristics are important for different job components and thus different jobs. Synthetic validation procedures should be able to make use of this variation where it exists. SYNVAL researchers (Peterson, Wise, & Campbell, 1991) have noted that one reason why the synthetic approach did not show greater differential validity could be that the jobs they examined (entry-level skilled positions) are not widely different in terms of their basic job requirements. In this regard, the degree of differential validity captured by the empirically based operations, while significant, was not overly large.

Most applications of the synthetic validation paradigm have used expert judgments to establish the relationships between human characteristics and job components. Although the research discussed here and other studies (e.g., Schmidt, Hunter, Croll, & McKenzie, 1983) have shown that these judgments can be reliable and reasonably equivalent to empirical validity coefficients, it should be noted that validity judgments may be affected by prior empirical results. Because trained psychologists are often used as raters, judgments of validities may be dependent upon past validity studies. This influence may not be a problem for most applications of synthetic validity; however, in situations where new predictor or criterion measures are being applied, it is possible that judgments between variables will be more difficult. More research is needed on the factors that may affect these judgments.

So far, the results of applied synthetic validation efforts have been encouraging. Future research should continue to examine the issues of discriminant validity under conditions where true validities are more likely to vary and the judgment process behind attribute requirement estimation. The other procedures for generalizing validities that are reviewed here are less dependent on human judgments.

Validity Generalization

In recent years, procedures have been developed for aggregating the results of multiple studies to develop a more statistically reliable and accurate estimate of the relationships among variables. These procedures have been applied generally, to the summarization of research findings in the form of a meta-analysis (Glass, McGaw, & Smith, 1981), and adapted specifically for the estimation of the relationship between individual characteristics and job performance (Hunter, Schmidt, & Jackson, 1982). Validity generalization analyses are based on the idea that any single estimate of validity is a function of both the true correlation between two variables as well as both systematic and unsystematic sources of error. Thus, a number of validity coefficients based on the same underlying true correlation would be expected to vary due to various kinds of error in local validity studies. This error is attributable to factors such as sampling error, variations in criterion reliability, and variations in restriction of range. By correcting for these errors it is possible to more accurately estimate the true relationship between the variables.

Validity Generalization Procedures

In terms of correcting for various sources of error variance, validity generalization procedures go beyond those meta-analysis procedures which are only concerned with the effects of sampling error. Hunter et al. (1982) and Schmidt (1988) have described the procedures for conducting a validity generalization analysis; a summary of these procedures is provided below.

First, studies that investigate a relationship of interest (e.g., the relationship between cognitive ability and job performance) are gathered, and the effect sizes from these studies are put on a common metric, such as a correlation coefficient. Second, the variance of this distribution of validity coefficients is corrected for "artifacts;" that is, for sources of variation that are properly labeled as sources of error. The three sources of error variance in estimates of r_{xy} that have received the most attention are sampling error, differences in criterion reliability across studies and differences in the degree of range restriction across studies. (Note that other artifacts, such as predictor unreliability, are also often corrected.) After subtracting the variance due to artifacts, the variance remaining in the distribution of coefficients can be characterized as an initial estimate of variance in validity estimates that is due to substantive differences across studies. Third, the mean of the distribution is then computed and corrected for mean effect of range restriction and the mean attenuating effect of unreliability in the criterion measures. Where it is possible, these corrections are made for each study before the average effect size across studies is calculated.

The mean and variability of the distribution of effect sizes are then examined in two ways to determine whether the validities "generalize." The distribution of validity coefficients can be examined to determine whether a large portion of the observed coefficients lie above a minimally useful level. If, for example, 95% of the coefficients lie above the useful level, it may be concluded that the construct assessed is useful across situations (i.e., validities can be said to generalize). Additionally, a stronger conclusion may be supported if a substantial portion (75% has been the rule of thumb) of the variability in the coefficients can be accounted for by artifacts. If this is the case, it may then be concluded that the corrected mean (r) of the effect size distribution is the best estimate of the true correlation (i.e., the population parameter) between the predictor and criterion (that is, situational specificity of validity is rejected).

If it is not possible to reject the hypothesis that the variance in effect sizes is greater than that expected due to artifact, then meaningful situational moderators of the remaining variability should be examined. If characteristics of the studies used in the analysis (such as the cognitive complexity of the jobs studied) are related to effect sizes, then those characteristics may be said to moderate validity. This step in the validity generalization procedures differs from other meta-analytic techniques in the sense that typical meta-analyses will search for moderators before accepting a generalizability hypothesis (Glass et al., 1981). Validity generalization procedures initiate a search for moderators only if substantial variability in effect sizes remains after accounting for artifactual variables.

These procedures address two critical hypotheses about the validity relationship: whether the validities generalize across situations, and whether situations affect true validity varies at all as a function of situation differences. It is important to note that these hypotheses are not mutually exclusive. That is, it is possible for validities to generalize (have a non-trivial value across situations), and yet show some variability due to situations (e.g., jobs). Some tests may simply have lower (yet still useful) validities for some types of jobs than for others. The second hypothesis is a stronger version of the validity generalization argument (i.e., true validity does not vary across situations) and is more difficult to support.

Research on validity generalization has produced some refinements in the procedures and suggests improvements in the theory. These issues will be examined in the following sections.

Issues in Validity Generalization

A number of issues are currently being debated with regard to validity generalization procedures. Many of these issues deal with the technical details of the formulas used to correct the variability in effect sizes and the estimates of artifactual bias that are used in the corrections (see Hartigan & Wigdor, 1989, for a concise review). Other issues are more substantive, however. These issues deal with the probability of falsely rejecting the situational specificity hypothesis, the inclusion of flawed research in the group of studies examined, and the precision of the definition of the variables being examined in the generalization hypothesis (Guion, 1991).

Power. The power to detect true situational moderation using the Schmidt and Hunter validity generalization procedures has been criticized on several fronts (e.g., Callender & Osburn, 1980, 1981; James, Demaree, & Mulaik, 1986). Callender, Osburn, and their colleagues (Callender & Osburn, 1980, 1981; Osburn, Callender, Greener, & Ashworth, 1983) have developed an alternative approach to validity generalization. In their model, Callender and Osburn test the situational specificity hypothesis using Monte Carlo methods to generate a distribution of validity coefficients where the true variability in the coefficients is known. These generated distributions have been used to test the power of the Schmidt and Hunter decision rule for rejecting the situational specificity hypothesis (Osburn et al., 1983). (Recall that Schmidt and Hunter suggest that if artifacts account for at least 75% of the observed variability in validity coefficients, validities are assumed to be based on a common population value.) It was found that both the Monte Carlo procedures and the 75% rule lacked sufficient power to detect low to moderate levels of true variance in validity coefficients when the sample sizes of the studies used were below 100. Whether the effect sizes (i.e. residual variance in the prior distribution) that yield low power are too small to be of much concern is a matter of judgment.

Research included. A second issue deals with whether a representative sample of studies has been included in any particular validity generalization analysis. Specifically, it is possible that many validity studies that did not find a significant relationship were not published and thus were not included in the validity generalization analysis (cf. Rosenthal, 1979). If a large number of unpublished studies that show non-significant results were not included in a generalization analysis, it is possible that the mean of the obtained validities will be upwardly biased. However, studies go unpublished for a variety of reasons, not the least of which is that they may be poorly designed. Thus, it may be argued that the unpublished studies that are not included in the analysis are excluded with good reason -- to include them would add unnecessary noise to the data. A middle ground on this issue may include problematic studies in the analysis where possible and investigate study quality as a moderator. This approach has been used in some meta-analyses (e.g., Gaugler, Rosenthal, Thornton, & Bentson, 1987).

Precision. A third issue relates to the precision with which the variables investigated in a validity generalization analysis are defined (Guion, 1991). In an effort to amass a large number of studies, and thus a large total N, researchers often include studies that use a variety of criterion measures. This practice may confuse the issue because different types of criterion measures may tap different latent variables (McCloy, 1990) both in terms of the performance content represented and the performance determinants that are allowed to influence criterion variance. Also, although the predictor measures included may have a general factor in common (e.g., cognitive ability), it is often the case that the measures differ in overall factor structure. This is a critical concern for validity generalization because, if the relationship that is being generalized is ill-defined, it will be difficult to use results of these analyses to make predictions in situations where the relationship has not been investigated. Schmidt, Hunter, and Pearlman (1981) have made the claim that performance is best measured as a unitary construct for validation purposes. Recent work explicating the latent structure of performance, however, suggests that understanding the dimensions of performance may enhance our ability to predict (e.g., Campbell, 1990).

It is little wonder that one of the most pervasive validity generalization findings to date is that cognitive ability relates to overall job performance in most jobs (Schmidt & Hunter, 1977), given that only the most general predictor-criterion relationships are sought. As validity generalization models continue to develop, it will be important to refine the definitions of the constructs that we seek to relate and better specify the relationships we seek to generalize.

One Parameter or Many?

One could think of validity generalization analyses as a search for the relationships among latent variables (i.e., the true correlations). In this context it is reasonable to expect that more than one latent relationship may underlie a distribution of validity coefficients. Consider the following example: suppose two classes of predictor variables (e.g., cognitive ability and personality characteristics) were considered in a validity generalization analysis. In the same analysis "performance" could be construed as having two components: volitional behavior (a "will-do" component) and task proficiency (a "can-do" component). If each type of measure relates to the others to at least some degree, then the resulting distribution of coefficients could be best described by a model that assumes there to be four sub-populations of coefficients, not just one. Criticisms of the "75 percent rule" for rejecting the situational specificity hypothesis have indicated that the existence of such sub-populations (and their corresponding population parameters) are likely to go undetected in many validity generalization analyses (e.g., James et al., 1986).

Recent work on validity generalization models has developed additional procedures for testing the notion that more than one true correlation may underlie a distribution of validity coefficients. For example, Bayesian estimation procedures that allow for the testing of hypotheses that are based on more than one true population correlation have been applied to the validity generalization problem (e.g., Hedges, 1988; Thomas, 1990). The model proposed by Thomas (1990) provides estimates of (1) the number of population correlation coefficients that best fit a given distribution of sample coefficients, (2) the values of those population parameters, (3) the proportion of sample correlations that correspond to each of the parameters, and (4) the true variance among the population parameters. A similar model has been proposed by Hedges (1988). In that model, the likelihood that the true variance in population parameters equals zero is examined in light of the data included in the meta-analysis. In each of these models the possibility that some true variability may exist in the underlying population coefficients is acknowledged.

Conclusions Regarding Validity Generalization

Validity generalization procedures allow the relationships among latent constructs to be estimated. Although situational variables probably account for far less variance in validity coefficients than was once thought, current developments in validity generalization models leave room for true variability in validities to be identified. These procedures should help personnel researchers to better understand the relationships between predictor and criterion constructs. Another procedure for estimating validities in the absence of criterion data, multilevel regression, assumes that true validity varies across situations (jobs or job components), and it explicitly attempts to capture that variability.

Multilevel Regression

Multilevel regression (MLR) is a statistical procedure for developing modeling equations for nested data[1]. For example, the procedure has been used in educational research for examining the effectiveness of new instructional techniques, where different techniques are applied to students in different schools such that only one technique is used in a given school. Thus, some variability in effectiveness is due to differences between schools (e.g., Braun, 1989). This problem is handled in MLR by predicting the variability in the relationship between the treatment and the dependent variable with the characteristics of the schools in which the treatment was administered. The viability of the treatment in new schools may then be estimated by using the characteristics of that school to predict the expected treatment-performance relationship.

This modeling procedure has been adapted for predicting job performance for military jobs where criterion data have not been collected (Harris et al., 1991). The procedure is based on the assumptions that the true relationship between person characteristics and job performance varies between jobs, and that the variability in the relationship is related to the characteristics of the jobs. The logic of the procedure is similar to that of the educational example above: the relationship between person attributes and performance is determined for a set of jobs with known characteristics. Variability in the regression parameters for each person attribute is addressed by treating the parameters as a dependent variable and regressing them on the job characteristics in a second set of equations. The relationship between person attributes and performance for new jobs (i.e., the parameters for the first-level equation) can then be estimated by inserting the characteristics of those jobs in the second level equations. The following section describes the procedure as it was operationalized in the Linkage projects conducted for the Navy and the Office of the Assistant Secretary of Defense (OASD; Harris et al., 1991; McCloy et al., 1992).

The Multilevel Job Performance Model

To investigate the importance of quality of recruits in the Military Services, the Department of Defense (DoD) initiated the Job Performance Measurement/Enlistment Standards Project, a long term research effort to define job performance in the military and establish linkages between recruit characteristics and performance (DoD, 1987). As a part of that effort,

[1] This procedure is also known as Hierarchical Linear Modeling (Bryk & Raudenbush, 1987).

several new performance measures, including hands-on performance tests, were developed for a sample of jobs in each Service. The relation between recruit characteristics (such as cognitive ability) and job performance was estimated for those jobs. Unfortunately, performance tests were not available for all jobs, because the tests are expensive and time consuming to develop and administer. Thus, Harris et al. (1991) developed a multilevel model of job performance for obtaining performance predictions for jobs without hands-on tests.

The multilevel model relates characteristics of people to their subsequent job performance, and it explicitly accounts for variation in the relationship between individual characteristics and performance that is due to job characteristics. This approach is operationalized as a multilevel regression model with the following general structure:

$$P_{ij} = \alpha_j + \beta_j T_{ij} + \gamma_j O_{ij} + \varepsilon_{ij}.$$

In this model, the performance of person i in job j (P_{ij}) is modeled by a constant that is dependent on job j (α_j); a vector of regression coefficients that is dependent on job j (β_j) multiplied by the individual's score on a test of interest (T_{ij}); a vector of regression coefficients that is dependent on job j (γ_j) multiplied by other predictors of interest (O_{ij}); and an error term (ε_{ij}). Note that parameter O_{ij} is included here only to show that multiple predictors can be included in the model.

In the second level of the model, the regression coefficients obtained in the first model (e.g., β_j) serve as criteria and job characteristics are used as predictors. Specifically, the subscripted coefficients in the model indicate that it is possible for the contribution of the corresponding variable to vary across jobs, and it is assumed that this variability can be accounted for by the characteristics of the jobs. The multilevel regression approach assumes that these coefficients can be estimated using the following structure:

$$\alpha_j = \alpha + \pi_\alpha M_j + \delta_{\alpha j}\ ,$$
$$\beta_j = \beta + \pi_\beta M_j + \delta_{\beta j}\ , \text{ and}$$
$$\gamma_j = \gamma + \pi_\gamma M_j + \delta_{\gamma j}\ .$$

The job-related moderator variables, M_j, are characteristics of the jobs of interest. The coefficients in π describe the degree to which the variance in the job-specific parameters (α_j, β_j, and γ_j) is due to job characteristics, M_j. The M_j variables may differ for each parameter, and the δ's are random errors. This job performance model allows for the estimation of performance from individual characteristics, over a range of jobs, while accounting for the modifying effect of different job characteristics.

A multilevel job performance model was used to develop linkage prediction equations for entry level military jobs (cf. Harris et al., 1991; McCloy et al., 1992). The model used the Armed Forces Qualification Test (AFQT), an ASVAB Technical composite (made up of scores on the Auto and Shop Information, Mechanical Comprehension, and Electronics Information subtests), educational attainment (high school diploma graduate or non-high school graduate), and experience (total months of service) as individual attribute measures. Hands-on performance tests were used as criterion measures.

To include job characteristics in the model, it was necessary that all of the jobs of interest (including those without performance data) be described with the same job characteristic variables. The job-descriptive variables (M_j) used in the research were developed from an analysis of similar civilian jobs. Four component scores were used to describe each job: *Working with Things*, *Cognitive Complexity*, *Working Conditions*, and *Fine Motor Control*. These component scores were obtained for 925 military jobs having complete descriptive data. The scores were then used as the M_j variables in the multilevel regression model.

Deriving job-specific prediction equations was completed by first estimating an initial multilevel performance equation using all 24 jobs in the sample for which hands-on performance test data were available. Individual job equations were formed by inserting the appropriate component scores into the job-level (second level) equations and solving for the parameter estimates, which in turn are inserted in the individual level equation to solve for the performance prediction. The research found that the job components did account for some statistically significant variation in the regression parameters.

Evaluation of the Model

The multilevel model has been evaluated in two ways (McCloy et al., 1992). First, because the initial model was estimated using only 24 jobs, the sensitivity of the model to any one job was a concern; thus a sensitivity analysis was conducted to determine how the model parameters varied when individual jobs were excluded. Second, the validity of the job-level prediction equations for jobs that lack criterion data was investigated with a cross-validity analysis.

The sensitivity analysis was conducted by examining the change in the equation parameters when the initial equation is estimated on a smaller sample of jobs. This analysis was accomplished by holding out each one of the 24 jobs in the sample, and re-estimating 24 different performance models using reduced samples of 23 jobs each. The various parameter estimates were then compared across models. The results showed that individual-level parameters (i.e., those for AFQT, ASVAB technical composite, education, and experience) were quite stable. The parameters for the job-level parameters were somewhat less stable, as would be expected due to the small number of jobs in the model. Additionally, predicted performance based on the reduced (23 job) models did not vary meaningfully across the different models. This analysis suggests that the model is relatively insensitive to the specific jobs in the sample. However, the inclusion of more jobs would improve the stability of the job-level parameters.

Cross-validity analyses were conducted to examine how well performance is predicted for jobs without criterion data. This was accomplished by treating jobs with performance data as if they lacked performance data and then generating prediction equations for those jobs. Cross-validation of the performance model involved using each of the "reduced" (23-job) models to generate a job-specific equation for each of the respective holdout jobs, as if those jobs lacked criterion data. These equations were then compared to ordinary least squares (OLS) equations generated for each of the holdout jobs using the hands-on performance test data as criteria for those jobs. The OLS procedure provides optimal equations given the sample data, thus providing a ceiling value for the multiple correlation. The procedure was repeated for each of the 24 jobs in the sample. After correcting the OLS R^2 estimates for shrinkage, there was little difference between the OLS R^2 estimates and the estimates derived from the multilevel model (most R^2 differences were less than .02). Larger differences tended to be found only for jobs with smaller sample sizes. The cross-validity analyses suggest that the multilevel model provides fairly accurate job-level prediction equations for jobs that lack criterion data.

Conclusions Regarding Multilevel Regression

The multilevel regression approach is useful because it provides a method for accurately developing prediction equations for jobs that lack criterion data. The approach is also useful for another reason -- more resources can be devoted to the measurement of performance because performance data do not need to be collected on all jobs for which test validity needs to be established. In the example discussed here, hands-on performance tests, which are difficult and costly to develop, were used as criteria for a small set of jobs. If the M_j variables account for relevant variance in the predictor-criterion relationship, the multilevel approach allows for the generation of prediction equations for other jobs where such involved performance measures are not available. Of course, the viability of the procedure depends on the degree of relevant variance in the regression parameters accounted for by the M_j variables and on the extent to which the sample of jobs having criterion data is representative of the population of jobs for which prediction equations are to be developed.

As researchers continue to define a criterion taxonomy and develop more comprehensive criterion measures, multilevel models may prove useful for generating initial prediction equations using jobs where extensive criterion measurement has been conducted. Multilevel modeling may also benefit from recent advances in validity generalization research. Specifically, as we become better able to specify the conditions under which true validity might be expected to vary, robust job characteristics may be better identified for defining the second-level job equations. Research in this area may benefit from an investigation of differential validity resulting from the various job-level equations, similar to that performed in the SYNVAL project.

Multilevel regression makes great demands of the data; and, to estimate the initial model, a sample of several jobs, with several individuals in each job, is required. The research described here used 24 jobs and a total N of 8464, and including more jobs would have improved the stability of the job-level parameters. However, as the taxonomies of person characteristics and performance become better defined and moderators of validity are identified, multilevel regression may become increasingly useful for modeling the relationships between predictors, criteria, and job characteristics.

General Conclusions

This chapter examined three procedures for estimating the validity of individual characteristics for predicting performance on jobs where criterion data are not available. In the introduction, a general framework was discussed that might describe the relationships between person characteristics, job characteristics, and performance components. Realistically, however, it is likely that the levels of uncertainty and dynamaticity present in the variables we study may make the development of such a framework an illusive goal. Nevertheless, the logic of such a framework (e.g., that finite taxonomies of person, job, and performance components can be developed and general laws can be established that describe the relations between these variables) has been applied in each of the areas that was described. The three procedures examined here each provide a methodology for generalizing criterion-related test validities. However, these procedures are only effective when they are based on valid theories of the predictor, criterion, and job characteristics domains.

A number of general conclusions follow from the research that was presented:

- Synthetic validation procedures have been shown to result in validation equations that are similar to empirically derived equations, but more research is needed on the judgments that are used in the procedures, particularly with regard to their ability to capture the true extent of differential validity that might exist.

- Much of the situational specificity that was once believed to moderate test validities is due to various artifacts associated with conducting local validation studies. More precise estimation of the "true variance" in validity coefficients will be in large measure a function of the descriptive power provided by substantive models of the latent structure of the predictor-performance space.

- Recent developments in models of validity generalization provide more powerful methods of detecting true situational (job) moderation of validities. These developments could help researchers to better isolate true variation in validity from variation due to error.

- A relatively new procedure, multilevel regression, provides another method for validating jobs that do not have criterion data. The procedure explicitly incorporates job characteristic information into a prediction model to account for any moderating effect that these characteristics may have.

Taken together, the procedures discussed here suggest one additional conclusion. That is, any comprehensive effort to establish validity in a generally applicable manner should jointly recognize the interrelationships among all three of the taxonomic areas discussed in the introduction (predictors, performance, and job characteristics). Note that the term "validity" in this context may refer to three distinct activities: (1) the identification of appropriate individual difference variables to be included in a prediction model, (2) the estimation of the criterion-related validity of those variables for given jobs, and (3) the estimation of the differential validity of those variables across jobs.

V. METHODS FOR SETTING STANDARDS ON PREDICTOR AND CRITERION MEASURES

Rodney A. McCloy

The terms "performance standards" and "selection standards" appear frequently in the personnel research literature and human resource management professional literature, as well as in the popular press. Performance standards refer to scores of special interest on the criterion and selection standards refer to scores of special interest on the predictor(s). The scores designated as standards are given special importance because individuals who fall above or below the standard are treated differently, or at least evaluated differently, regardless of the scores earned by other individuals. This is the distinction of norm referenced versus criterion referenced measurement, as it is made in educational measurement. For performance standards, some "standards" of interest are those which distinguish: a) needs training vs. does not need training, b) should be fired vs. should not be fired, c) should be promoted vs. should not be promoted, or d) should hire people who would perform at this level or above vs. should not hire people who would perform below this level. For selection and classification the standards of interest are the predictor scores which determine hire versus not hire and the scores which govern entry into specific jobs during job assignment. While the terms selection standard and performance standard are familiar, identifying such scores via a reliable and valid scaling procedure is quite another matter. It has proven to be a very difficult measurement problem, with no consensus about how it should be accomplished.

Numerous methods for setting test standards have been proposed. In this chapter, we present a brief review of the more common methods and discuss the results of empirical studies that have compared them. This discussion is followed by a description of a somewhat different approach that goes well beyond the issue of setting a single "cut score" to modeling the variables affected when a standard goes operational.

Standard Setting Methods

Methods for setting standards on performance or predictor measures can be divided into two classes: item-based methods and examinee-based methods. Although most of the standard setting literature springs from educational research where the primary concern is setting competence on written multiple choice tests, the methods can be modified with varying degrees of success for use with performance tests (Jaeger & Keller-McNulty, 1991).

Item-Based Methods

Item-based methods require raters to make judgments regarding the proportion of minimally competent individuals (i.e. minimally competent vs. non minimally competent = "the standard") who would correctly answer each test item. These methods are more widely used than examinee-based methods.

Angoff method. For the Angoff (1971) method, judges are instructed to estimate the percentage of minimally competent individuals who would correctly answer each test item. The percentage estimated to pass each item is converted to the percentage of items that should be passed by minimally competent individuals.

The Angoff method is arguably the easiest method to implement. Judges have little problem understanding their task. The method has its drawbacks, however, as do all the methods to be described. For example, the method often results in highly variable standards across judges unless coupled with normative data or an iterative procedure (Jaeger & Busch, 1984; Norcini, Lipner, & Langdon, 1987). In addition, the method is appropriate only for dichotomous items. The method could be modified for continuous measures by asking subject matter experts (SMEs) to estimate the most likely, or average, score for minimally competent individuals.

Nedelsky method. The Nedelsky (1954) method requires a multiple choice format. Judges must identify the distractors a minimally competent individual would readily eliminate as incorrect. A minimum passing level (MPL) is then calculated for each item by taking the reciprocal of the number of response options a minimally competent individual could not identify as incorrect. A standard for each judge is obtained by summing the judge's MPLs for the test items. The standard for the test is the average of the judges' standards.

The Nedelsky method suffers several drawbacks. For example, it is assumed that examinees randomly select a response option from those that cannot be identified as incorrect. This assumption depends upon a second assumption that examinees do not draw on partial information from the item or its distractors. The effect of these assumptions is a standard that is more lenient than that obtained using other methods.

Perhaps more damaging to the method is its difficulty. Judges often report being confused and having little confidence in their judgments (Poggio, 1984). Finally, the method requires a multiple choice test. Although appropriate for written tests of job knowledge, the Nedelsky method cannot be used with work samples or performance ratings.

Ebel method. This method, described by Ebel in 1972, also involves judgment about minimally competent examinees. Judges must classify items into the cells of a matrix defined by some number of levels of difficulty and relevance. Although the number of levels of difficulty and relevance can vary, Ebel suggest three for difficulty (easy, medium, and hard) and four levels of relevance (essential, important, acceptable, and questionable). Working together, the judges estimate the percentage of minimally competent individuals who would correctly answer a large sample of items similar to the items in each cell.

Unlike the Nedelsky method, judges readily understand this judgement task. The method does possess drawbacks, however, such as its tendency to be quite time-consuming and to result in consistently stricter standards than other methods (Andrew & Hecht, 1976; Poggio, 1984; Skakun & Kling, 1980). But its major drawback for use in the military is that few military performance measures are commensurate with the categorization the method requires (Jaeger & Keller-McNulty, 1991). If the dimensions are ill-defined, the standards are like to be highly variable and unstable.

Jaeger method. Rather than having judges estimate the percentage of minimally competent examinees who would answer an item correctly, Jaeger's (1982) method asks judges the following question: "Should *every* examinee in the population of those who receive favorable action on the decision that underlies use of the test (e.g., every enlistee who is admitted to the military occupational specialty) be able to answer the test item correctly?" (Jaeger & Keller-McNulty, 1991, p. 268). The method also employs an iterative approach coupled with normative data. Specifically, after answering the preceding question for each test item, judges as a group receive information on the percentage of examinees who actually did answer each item correctly during a recent administration of the test. After considering these data, judges reconsider their estimates and then independently respond to the same question for each item a second time. For the final phase of the judgement exercise, more normative data are provided, this time information on the number of examinees who would have failed the test had the judges' standards been adopted. Judges are then given one more opportunity to amend their estimates.

Given the group discussion format of the iterative procedure, the Jaeger method can be rather time-consuming. The method does have the distinct advantage of being applicable for any type of measure, although some modification would be required before applying it to continuously scored measures (e.g., ratings).

Although the possibility of judges defining their referent groups differently is not a problem, the Jaeger method does not rule out the potential for different standards to affect the judgments. For example, there could be conflicting ideas about the items that should be correctly answered by individuals who receive a favorable personnel action. Although the iterative procedure provides some buffer against such an occurrence, untoward group dynamics (e.g., a highly dominant or persuasive individual) could subvert the process.

Examinee-Based Methods

Methods based on examinees rest primarily on two assumptions. The first is that judges who are familiar with examinee performance in the area being tested can identify high- and low-performing individuals. A second assumption underlying examinee-based methods is that most judges are more comfortable making decisions about individuals than about test items.

Borderline-group method. This method determines a test standard based on actual borderline (i.e., minimally competent) examinees (Livingston & Zieky, 1982). Judges are asked to identify competent, borderline, and incompetent examinees. The cut score is the median score of the borderline group, but this value is often adjusted downward slightly to account for measurement error.

Contrasting-groups method. For this method, judges must identify two groups of examinees: those who they are sure demonstrate mastery of the material being tested, and those who they are sure do not demonstrate mastery. The test scores for the two groups are plotted and compared. The score appearing at the point of intersection for the two distributions is selected as the standard.

Estimates of time to proficiency as a means of standard setting. Another type of examinee-based method requires judges to estimate the amount of time it requires an individual to become proficient on the job. Rather than having supervisors identify groups of exceptional and marginal examinees, recent research for the Navy (Harris, personal communication, Sept., 1992) asks the supervisors to estimate the amount of time it takes an average recruit (i.e., AFQT[1] of 50) to attain acceptable performance in their rating. Estimates may also be obtained for the time required to demonstrate other levels of performance (e.g., minimal or exceptional levels of proficiency).

This method does have the severe drawback that supervisors have been shown to be unable to provide reliable ratings of time to proficiency (Leighton et al., 1992). More research is needed to examine the possibility that such ratings could be made more reliably if supervisors are (1) given the opportunity to observe key tasks and (2) instructed to pay particular attention to the key tasks for purposes of performance assessment.

Summary. The standard setting methods described above are the most commonly applied procedures, but there are others (e.g., Berk, 1976; Kriewall, 1972). Comparison of the methods by empirical analyses have generally found that the Ebel method produces the strictest standards, whereas the Nedelsky method produces the lowest standards. The standards resulting from the Angoff and Jaeger methods typically fall somewhere in between. Although all the methods have certain disadvantages, Berk (1986) argued that the Angoff method is the best in terms of technical adequacy and applicability. He also gave high marks to the contrasting groups method.

The most pervasive finding of the studies that have compared the various standard setting methods has been suggested in this section--namely, the different methods result in different standards (e.g., Andrew & Hecht, 1976; Sigmon & Halpin, 1984; Livingston & Zieky, 1983; Skakun & Kling, 1980). Because of the disparity in standards established by the various procedures, many researchers recommend the use of several standard setting procedures (Halpin et al., 1983; Koffler, 1980).

[1] The Armed Forces Qualification Test (AFQT) is a composite score comprising the verbal and mathematical subtests of the Armed Services Vocational Aptitude Battery (ASVAB), the examination administered to military applicants.

Similar results were obtained by Wise, Peterson, Hoffman, Campbell, and Arabian (1991) during the Army's Synthetic Validation Project. These researchers used two standard setting methods not described above. In the behavioral incident method, judges read a description of a specific incident of job performance. The judges were instructed to consider a soldier performing similarly on a consistent basis and to rate that soldier's performance as unacceptable, marginal, acceptable, or outstanding. The second method, a task-based standard setting procedure, presented judges with a form describing three sample tasks from the Project A hands-on tests. A distribution of the hands-on test scores was also provided for the dimension in question, along with values indicating the percent of soldiers who scored at or below that particular test score in the Project A database. Judges were to indicate cut scores on the distribution by drawing three lines to demarcate unacceptable, marginal, acceptable, and outstanding performance on the dimension.

As with previous research comparing standard setting methods, the standards generated from these two methods were quite different:

> "As many as half of current incumbents were less than fully acceptable and nearly 30 percent were unacceptable according to the standards set by the task-based method. By contrast, fewer than 40 percent of current incumbents were less than fully acceptable and only 6 percent were unacceptable according to the standards set by the behavioral incident method" (Wise et al., 1991, p. 7-7).

Although the method chosen has a substantial impact upon the standards derived, other factors also influence the standards. Pulakos, Wise, Arabian, Heon, and Delaplane (1989) presented a model of the standard setting process that relates such factors as the characteristics of the judges (e.g., demographics, knowledge, interest group), the number of judges, the type and amount of training judges receive, the choice of judgment facilitation techniques (e.g., normative data), and the purpose of the research (e.g., the number of standards to be set, the use of the standards), and so on.

Cut Scores, Distributions, and Costs

Most of the work on standard setting methods has been done in education where the goal typically has been to give operational meaning to particular levels of competence or mastery (e.g., he knows enough to go on to the next course) that do not require comparisons to other people (i.e., are not norm-referenced). To make this type of interpretation, one must be able to assign a meaning to the test score itself, such as what the score implies about how much of the content domain the individual has mastered. Testing of this sort has been labelled criterion-referenced or domain-referenced testing.

Organizations like the Services need the same thing when they must make real-time selection/classification decisions and cannot use top-down selection or assign people to jobs in large batches such that some objective function is maximized. They also may not want to fill all available slots if the predicted performance for a individual is below some critical standard. Selection "standards" only have meaning in terms of their relationship to performance standards. The criterion-referenced scaling methods just discussed apply directly to the problem of setting job performance standards, but we know of no instances where the methods have been applied successfully such that the critical scores and the measurement operations on which they were based exhibited high reliability and at least a minimum level of construct validity. The usual outcome is a severe lack of consensus about what the "standard" should be. The lack of consensus about what performance should mean, what measures are valid reflections of it, and how high is high make standard setting a virtually impossible task, as suggested by the finding that different methods often yield markedly discrepant standards.

But even if all of the methods discussed provided the same standards, problems of interpretation would remain. One must consider the ramifications the standards will have for selection and classification. For example, one variable that has not been addressed in the preceding discussion that has enormous policy relevance is cost. With every standard, there is some cost associated with designating a group of individuals as marginal (e.g., remedial training costs) or unacceptable (e.g., recruiting and training costs to replace the individual if discharged from military service). Similarly, the recruiting cost of those exceeding a specified cut score will typically increase as the cut score increases, given the tendency for higher quality individuals to be more expensive to obtain (e.g., McCloy et al., 1992).

The classic economic solution says the organization should keep spending until the marginal revenue equals the marginal cost. Models for determining this point are termed cost/benefit models. Although defensible in the private sector's free market economy, there is no way to represent the benefit side adequately in the public sector. Although some have argued that public sector organizations purchase inputs in competitive public sector markets and thus pay the market price (e.g., Nord & Kearl, 1990), this does not imply that the value of the marginal product of those inputs in the public sector organization is equal to the input price (cf. McCloy et al., 1992, p. 16).

One way of avoiding the problem of measuring payoffs (hence, valuing performance) is to pose a different set of questions that can be addressed more easily. For example, one can ask: (1) For a given level of aggregate performance, or a given distribution of aggregate performance (in whatever metric), what personnel management strategy will minimize the cost of achieving it?, or (2) For a given amount of money to spend, what personnel management strategy will maximize aggregate performance? Models aimed at these questions are termed cost/effectiveness models. Note, however, that emphasis is shifted to getting the most performance from a given investment or to minimizing the cost of maintaining performance at its current levels.

Cost Effectiveness and Cost/Performance Tradeoff Models for Evaluating Selection and Classification Standards

Prior to the Joint-Service Job Performance Measurement/Enlistment Standards (JPM) Project which was begun by the Department of Defense in 1980, military job performance measures provided normative information only (i.e., what the relative standing of individuals is in a given job). Green and Wigdor (1988) called for the development of domain-referenced performance measures that would allow performance scores to be interpreted in terms of competence (cf. Mayberry, 1987). They then stated that in selection, cut scores (i.e., standards) are placed on the selection tests, not on the performance measures they are designed to predict. As such, policy makers should examine the performance distribution that results from the use of a particular predictor cut score. Evaluation of the cut score will follow from the assessment of whether the obtained performance distribution is desirable. Such a focus (1) again emphasizes that selection cut scores have meaning only in terms of their relationship to performance standards, and (2) stresses the entire range of obtained performance, going well beyond the simplistic notion of competence/incompetence. Green and Wigdor emphasized, however, that it is feasible a standard also could be placed on the performance measure, leading to consideration of the distribution of examinees on the selection test(s). Either way, the focus shifts from a single cut score to a distribution resulting from the cut score. The cut score may be altered until a desired distribution is obtained.

The distribution must be interpreted in terms of the competence of the individuals constituting it. Policy constraints must also be considered, including end-strength goals, attrition rates, and costs. The primary goal of the Enlistment Standards portion of the JPM Project was to establish the linkage (i.e., the relationship) between job performance and enlistment standards. Several models considering these variables and a distribution of performance or recruit quality have been developed.

The Armor Model. Perhaps the most prominent model in the literature is the one developed by Armor, Fernandez, Bers, and Schwarzbach (1982) and Fernandez and Garfinkle (1985). Armor and his colleagues described a cost-effectiveness model that solves for an AFQT cut score that minimizes the cost of achieving a given level of first term performance by recruits entering a given occupation. The measure of performance used in the Armor model is the expected number of "qualified" first term man years, where "qualified" was determined by the probability that an applicant will "pass" the job-specific Army Skill Qualification Test (SQT). Passing was simply defined as a percent correct score of 70. The probability of passing is estimated from a model relating the entry characteristics of soldiers in a given Army occupation to the scores those soldiers obtain on the SQT. Hence, the output measure combines the probability that an applicant will survive to a given point in the first term of Army service with the probability that the applicant will have passed the SQT to obtain a measure of "qualified man years."

The cut score chosen by the Armor model is the one that minimizes the recruiting, training, and compensation costs over the first term of service, subject to letting sufficient numbers of applicants enter so that a specified level of expected qualified man years of service (i.e., a performance goal) is achieved.

The Nord and Kearl Model. Nord and Kearl (1990) incorporated the Schmidt and Hunter estimate of the value of the standard deviation of performance in dollars (i.e. SDy = 40%

of average salary) in an optimization model for determining recruit quality mix, thus developing a cost-benefit model (as opposed to a cost-effectiveness model). They use this estimate in an optimization process to solve for the level of performance at which the marginal value is just equal to the marginal cost. Note that the solution depends upon the validity of representing the value of different levels of soldier performance in terms of a dollar return to the organization. The model attempts to solve for the optimal numbers of each of several categories of recruits in each of several discrete recruit categories, considering all occupations simultaneously. The Armor model, on the other hand, solves for an optimal "cut" score--a minimum score recruits must achieve to enter the Armed Forces. Although this cut score approach is a pragmatic solution to the entry standards problem, it can result in a solution that is at best the same, and is generally more costly, for a given performance goal than the solution for a model solving for the optimal number of recruits from discrete quality categories. In addition, the Armor model considers only one occupation at a time. Although this greatly simplifies the optimization problem, the recruiting costs used to solve for the optimal cutoff score are unlikely to be correct. The problem is that this approach fails to consider the demand for high quality recruits in other occupations when solving for the cutoff score of the occupation being analyzed.

Cost/Performance Tradeoff Model. Building on this research, McCloy et al. (1992) described a model that solves for the most cost-effective recruit quality mix (quality being defined as a function of AFQT category and high school graduate status) meeting specified performance goals across occupations. The model therefore implicitly selects and classifies recruits at once, taking into account relevant costs and comparative advantage.

The objective function for the cost/performance tradeoff model is to choose recruits from each of several recruit quality categories to enter various military occupational categories, so that the sum of recruiting costs, training costs, and compensation costs is minimized, subject to meeting the performance goals in each military occupation, and, optionally, accession or first term end strength constraints (Hogan & Smith, 1991). As such, it is well suited to an analysis of the effects of changes in performance goals, strength, or accession constraints on costs and optimal recruit quality goals. The effects of budget changes on the optimal recruit quality mix and on performance must be analyzed indirectly however, because costs are the quantity that is minimized in the objective function. McCloy et al. (1992) discussed ways of addressing this dual minimization problem.

The cost/performance tradeoff model comprises four primary components: (1) linkage (i.e., prediction) equations, describing the relationship between individual characteristics (AFQT score, ASVAB Technical composite score, high school graduate status, and time in military service) and job performance (operationalized as one's score on a job-specific hands-on performance test), (2) survival rates for each occupation, (3) a recruiting cost function, and (4) training and compensation costs.

The model uses performance prediction equations developed from data generated by the JPM project. These prediction equations, estimated using a multilevel random coefficients model, allow generalization across all military occupations to predict a potential new entrant's performance in any military occupation, given the applicant's education level and AFQT category.

Survival rates were estimated directly from a historical cohort of accessions. These values represent the average number of months recruits remained in their job during the first term of service for each recruit quality category and occupation. The survival estimates are used to weight the predicted performance scores to give the expected number of man years of performance for recruits, again by quality category and occupation.

The recruiting cost function is derived from the estimated relationship between factors affecting recruiting, including recruiting resources (e.g., bonuses, number of recruiters), and the actual quantity of recruits. This function allows estimation of the recruiting cost for different numbers and quality mixes of recruits. The recruiting cost function is an optimization model itself, in that it estimates the least costly mix of recruiting resources as a function of the number and quality mix of recruits, the prices of recruiting resources (e.g., advertising), and other factors affecting the recruiting environment not controlled by policy (e.g., unemployment rate).

The costs of basic training and initial skill training are included.[2] Basic training is constant within a Service. Initial skill training varies by occupational category. Compensation costs include basic pay, allowances, and retirement accrual over the first term. The estimates are only as good as the validity of the functions that specify the effects of investments on recruiting and training results.

Using quadratic programming techniques, the model simultaneously selects and implicitly classifies recruits into one of 36 occupational categories, to minimize the recruiting, training, and compensation costs of meeting performance goals in those occupations over the first term of service. With the cost/performance tradeoff model, the performance goal serves as the standard on the performance test. The model solves for the cost-minimizing recruit quality mix (i.e., distribution) that will meet the standard in question. The optimization is influenced by important policy variables, such as constraints on the number of accessions and requirements of the number of high quality recruits that need to be assigned to each job.

Note, however, that the model does not solve for the "optimal" performance level of the first term force. This question of "how much performance is enough?" would require an appropriate utility metric to be placed on performance. Such valuation is very difficult in the private sector and of questionable meaning in the public sector, given that there is no market, *per se*, and reaping a profit is not a goal of the organizations in question. The most important point to make is that the model provides policy makers with information to which they may apply their own valuation criteria. The ramifications of their performance standards are quantified in the distributions of performance and recruit quality they yield. Further, the various components of

[2] Nuclear power training is also included for the Navy.

the model may be modified through "what if" analyses which can provide additional information regarding the dynamics of the selection system.

Standard Setting and the Cost/Performance Tradeoff Model. Standard setting requires difficult to make value judgments. The cost/performance tradeoff model does not make them less so, but it does provide the type of information that should make the consequences of alternative courses of action clearer. For example, recall that one of the variables in the performance equation used in the cost performance tradeoff model is time in service. If time-to-proficiency estimates are available, then the mean estimate of time to acceptable performance can be inserted into the performance equation, along with the conditional mean values for the other predictors, to generate a predicted hands-on performance test score for that time in service. Inserting values of one standard deviation above and below the mean time to acceptable performance into the corresponding job-specific performance equation provides a reasonable range of values on the hands-on test. To account for measurement error, the lower bound could be selected as the standard for acceptable performance in the rating. Further, the standards for all the ratings could then be used as performance goals in the cost/performance tradeoff model and the cost implications for these performance goals evaluated through the solution of the most cost-effective recruit quality mix that would meet the goals, as described above.

It should be noted that the tradeoff model does not suggest cutoff scores that should be placed on aptitude area composite scores for each job in the military. In this sense, the model cannot be used for setting job-level selection or classification standards. Rather, it can be used, as just described, to evaluate the costs of setting various alternative goals for performance or recruit quality. If, however, one is willing to describe "standards" as the desired recruit quality mix, the model does provide plenty of relevant, useful information regarding the cost and performance ramifications of particular enlistment standards.

Summary

Rarely does one encounter a concept that is seemingly so straightforward and yet in truth is so complex as setting standards on predictor or performance measures. Clearly, there is no such thing as a correct standard in the absolute. Some models, such as the cost/performance tradeoff model described above, can solve for optimal standards given certain constraints (e.g., the maximal level of performance that can be obtained given a certain increase or decrease in the recruiting budget), but these standards would change as soon as other variables in the model changed. The numerous methods described in this chapter further speak to the notion that a standard may have utility for a particular purpose (e.g., setting a minimum score for licensing veterinarians), but it can only be evaluated with respect to the results of its application (e.g., Are we satisfied with the quality of the licensed veterinarians? Are we experiencing a shortage of veterinarians because the standard is too high?). Weitz (1961) once provided a list of criteria that could be used for evaluating criteria. Perhaps by considering distributions of performance or recruit quality and the costs associated with the implementation of standards, an acceptable list of standards for standards can be developed. Such progress at least suggests that our competence in the area of standard setting might be increasing.

VI. MODELS OF FAIRNESS

Teresa L. Russell

What is Fairness?

Fairness, regardless of context, can be an elusive concept. With regard to employment decision-making, definitions of fairness must take into account societal notions about what is fair and the organization's values, as well as the ramifications of different procedures for the productivity of the organization. Attempts to define fairness without explicating the subjective, value-laden components muddle societal and psychometric goals. Such definitions can appear disingenuous or lack internal consistency. With this in mind, the Society of Industrial and Organizational Psychologists (1987) defined fairness as a social rather than a psychometric concept. Fairness, like validity, is a function of how test scores are used for the job and the population at hand.

Even so, fairness must be defined operationally to evaluate employee selection procedures. Such attempts have focused on the content, psychometric properties, and use of tests as well as the outcomes of testing. For example, Helms (1992) recently argued that traditional cognitive tests are unfair, based on their content. She asserted that cognitive ability tests measure attributes that are defined by Eurocentric values. Fair tests would measure cognitive attributes defined by other cultures or in the language of other cultures. Her jeremiad is reminiscent of that which led the attempts to develop culture fair tests in the 1950s, 60s, and 70s. Most of those studies found that race differences evidenced on so-called culture fair tests were in the same direction as those on traditional measures (Jensen [1980] for a discussion of culture fair test data). Moreover, definitions of bias in terms of test content have not explained large differences in test scores.

Adverse impact, on the other hand, is an outcome-based definition of fairness. It occurs when there is "a substantially different rate of selection in hiring, promotion, or other employment decision that works to the disadvantage of members of a race, sex, or ethnic group" (American Institutes for Research, 1992). Adverse impact is not, however, proof of unfairness.

Cleary's (1968) psychometric model of fairness is currently accepted by both the <u>Uniform Guidelines</u> (1978) and the Society for Industrial and Organizational Psychology (SIOP, 1987). The Cleary model distinguishes between test bias and test fairness: "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup" (Cleary, 1968, p. 115). In other words, a test is biased when prediction from a common regression equation results in either over- or underprediction of subgroup performance; this is called differential prediction. Overprediction of the performance of a protected group, when a common regression line is used, indicates bias but is generally not considered a fairness problem (SIOP, 1987).

It is important to distinguish differential validity from differential prediction. Differential validity occurs when the observed validity coefficient for one group is significantly different from the observed validity for the second group.[1] Literature reviews have concluded that differential validity occurs rarely and when it does occur differences between validity coefficients for blacks and whites are generally small (Cascio, 1982; Hunter, Schmidt, & Hunter, 1979; Linn, 1978). More important, differential validity alone would not be a sufficient indicator of unfairness even if it were a common finding (Bobko & Bartlett, 1978; Linn, 1978). "Different validity coefficients can occur for two groups even though the groups have identical prediction systems simply because one group has a greater variability..." (Linn, 1975, p. 298). Moreover, differences in prediction systems (e.g., slopes, intercepts, standard errors of estimate) are more directly related to issues of bias in selection than are differences in correlations.

Based on a review of differential prediction research, SIOP (1987) concluded that "there is little evidence to suggest that there is differential prediction for the sexes, and the literature indicates that differential prediction on the basis of cognitive tests is not supported for the major ethnic groups" (p. 18).

Other events suggest that differential prediction and fairness issues are not so easily discarded. Linn (in press) reviewed several studies citing evidence of differential prediction (e.g., Houston & Novick, 1987; Dunbar & Novick, 1988) and suggested that SIOP's dismissal of differential prediction was premature. Also, the National Academy of Sciences (NAS) chose to deviate from the Cleary model in making recommendations for the use of General Aptitude Battery Test (GATB) scores (Hartigan & Wigdor, 1989), and the GATB report left several questions regarding the fair use of tests unanswered for the business community (cf. "More normal nonsense," 1989). Soon after publication of the GATB report, the Civil Rights Act (CRA) of 1991 prohibited adjustment or use of scores on the basis of group membership, a procedure implied by several fairness models. Specifically, Title 1 Section 106 of the CRA of 1991 states:

> It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of the race, color, religion, sex, or national origin.

Consequently, there is uncertainty about the current operational definition of fairness.

[1] There was much debate, and confusion, about the definitions of differential prediction, differential validity, and single group validity in the 1970s. Definitions involved testing various combinations of hypotheses about whether the validity coefficients for two groups were significantly different from zero and/or each other. For example, Boehm (1972) defined single group validity as the situation where (a) the observed validity coefficient is significantly greater than zero in one group but not the other and (b) there is no significant difference between the two observed correlations. But Bartlett, Bobko, and Pine (1977) pointed out that this occurrence is a sample outcome; single-group validity cannot exist in the population.

The goal of this chapter is to examine fairness issues within both a psychometric and societal context. It begins with a short review of psychometric fairness models and moves toward current societal, pragmatic, and research issues in later sections.

Models of Fairness

Several major models of fairness were proposed in the 1970s--the regression model, the constant ratio model, the conditional probability model, and the equal risk model. Few new ideas have been proposed other than the structural equity model (Gregory, 1991)--a way of defining fairness--and score banding (Cascio, Outtz, Zedeck, & Goldstein, 1991)--a way of using test scores.

Cleary (1968) Regression Model

As mentioned, according to the Cleary (1968) model a test is biased when prediction from a common regression equation results in either over- or underprediction of subgroup performance. To illustrate her method, Cleary compared regression equations for the prediction of college grades from Scholastic Aptitude Test (SAT) scores for Black and White students from three schools. There were no significant differences between regression lines for Black and White students in two schools; however, Black students' scores were overpredicted by use of the white or common regression line in the third school.

Cleary's definition is a straightforward application of least squares regression. It is, therefore, an optimal definition with high expected utility (Hunter, Schmidt, & Rauschenberger, 1977). Early on, some critics suggested that because tests are not perfectly reliable, Cleary's model would disguise bias against the lower scoring group. But in turn, Hunter and Schmidt (1976) showed that an unreliable test is biased against the better qualified applicants (i.e., those with the higher intercept).

True application of Cleary's method is to use the common regression line if there is no differential prediction and to use separate regression lines if differential prediction occurs.

Thorndike (1971) Constant Ratio Model

Thorndike (1971) stressed that group differences on the criterion as well as the predictor must be accounted for in the regression model. He suggested that "qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance" (p. 63). Here, a selection measure is fair if the ratio of the proportion selected to the proportion successful is the same in all subpopulations (Cole, 1973). For example, a procedure would be considered fair if (a) 20 percent of Group A were selected while 80 percent of Group A was expected to perform successfully (a ratio of 1:4) and (b) 10 percent of Group B were selected while 40 percent of Group B was expected to perform successfully (a ratio of 1:4). Thus, passing scores on the predictor are set at levels that qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance.

Although Thorndike's definition generally yields societally appealing results (Petersen & Novick, 1976), several authors have expressed concerns about the constant ratio model. Hunter and Schmidt (1976) argued that Thorndike's method was essentially quota setting and expressed concern about setting quotas on the basis of sample data. Hunter et al. (1977) pointed out that overprediction of minority group performance is fair according to Thorndike's definition. That is, Thorndike's definition requires that the standard score difference between subgroups on the test be equal to the standard score difference on the criterion (corrected for unreliability). "Unless validity is perfect or unless there are no subgroup differences on the criterion, this requires the majority regression line to lie above the minority regression line, that is, it requires the test to overpredict performance for the minority subgroup" (p. 245).

Petersen and Novick (1976) determined that the Thorndike model is logically inconsistent. They examined the internal consistency of several fairness models by considering the converse of each model. The constant ratio model was restated to define a procedure as fair if the proportion of applicants rejected to the proportion unsuccessful is the same in all subpopulations. The selection and rejection specifications could only be simultaneously satisfied when the ratio was 1:1 or when the probability of successful performance was the same in both groups.

Cole (1973) Conditional Probability Model

According to the conditional probability model, a test is fair if the ratio of individuals selected to individuals who could, if selected, perform successfully is the same across subgroups:

> The basic principle of the conditional probability selection model is that for both minority and majority groups whose members can achieve a satisfactory criterion score [$Y > Y_p$] there should be the same probability of selection regardless of group membership (Cole, 1973, p. 240).

Petersen and Novick (1976) illustrated the difference between Thorndike's constant ratio model and the conditional probability model in terms of a bivariate distribution (see Figure 6.1). The conditional probability model refers to the ratio of the number of applicants in Region I (true positives) compared to the number of applicants in Regions I and II combined (true positives and false negatives). The constant ratio model considers the ratio of the numbers of applicants in Regions I and IV (true positives and false positives) to the numbers of applicants in regions I and II (true positives and false negatives). Like the constant ratio model, the conditional probability model is consistent with its converse (i.e., equal conditional probability of rejection given failure) only under very specific conditions (Petersen & Novick, 1976).

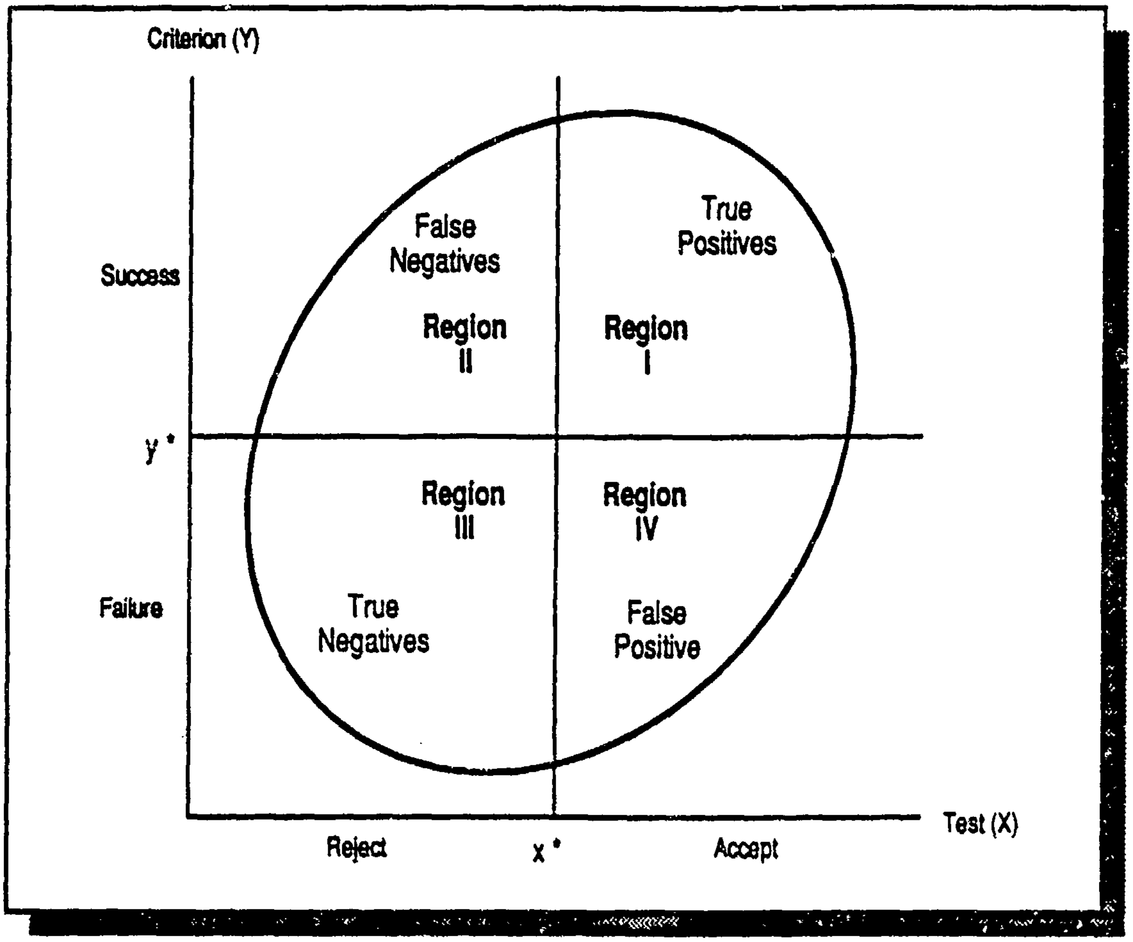

Figure 6.1 A Hypothetical Bivariate Distribution

Note: From "An evaluation of some models for culture-fair selection" by N.S. Petersen and M.R. Novick, 1976, in Journal of Educational Measurements, 13, p.10

Some data from a GATB validity study on carpenters also illustrates the difference between these two models (cf. Hartigan & Wigdor, 1989, p. 198). The study included 91 white and 45 black job incumbents. Table 6.1 shows frequency counts--the numbers of whites and blacks according to test performance (pass/fail) and job performance (good/poor).

In this example, the test overpredicted black job performance. Fifty percent of white carpenters (11 of 22) who failed the test performed well on the job (i.e., were false negatives), compared to 25 percent of black carpenters in the false negative category (8 of 32). Thirteen percent of whites (9 of 69) and 38 percent of blacks (5 of 13) fell into the false positive region. In sum, a higher percentage of whites than blacks who failed the test would have been satisfactory workers, and a higher percentage of blacks than whites passed the test but performed poorly.

Table 6.1
Numbers of Blacks and Whites According to Test and Job Performance

	Test Performance					
	Whites (N=91)			Blacks (N=45)		
Job Performance	**Fail**	**Pass**	**Total**	**Fail**	**Pass**	**Total**
Good	11	60	71	8	8	16
Poor	11	9	20	24	5	29
Total	22	69	91	32	13	45

The Thorndike definition of fairness would lead to the conclusion that the test is biased against blacks. Sixty-nine whites passed the test, while 71 performed successfully (A ratio of 69:71, or 97:100). Thirteen blacks passed the test, and 16 performed successfully (13:16, or 81:100). The test is not fair because the ratio of the proportion selected to the proportion successful is not the same in all subpopulations.

Similarly, the Cole model would find test bias. Whites who could, if selected, perform successfully on the job were more likely than blacks to be selected. Sixty of 71 whites (i.e., 85 percent) were good workers who were also selected by the test. In comparison only 8 of 16 blacks (i.e., 50 percent) who performed successfully were selected by the test.

Darlington (1971) Subjective Regression Model

Darlington stated four definitions of fairness in correlational terms:

(1) $r_{CX} = r_{CY}/r_{XY}$
(2) $r_{CX} = r_{CY}$
(3) $r_{CX} = r_{CY}r_{XY}$
(4) $r_{CX} = 0$

where Y is the criterion variable, X is the predictor variable, and C denotes an applicant's cultural group membership. r_{CX} measures the degree to which the test discriminates among cultural groups; r_{CY} is the correlation between cultural group and the criterion, and r_{XY} is the test's validity. Definition (1) is equivalent to the regression model; it states that the test is fair if knowledge of a person's cultural group does not improve the prediction of Y made from X. Definition (2) is the same as Thorndike's constant ratio model. Definition (3), Darlington's preferred definition, is a special case of Cole's conditional probability model. Definition (4) states that a test should not correlate with culture, regardless of the values of r_{CY} or r_{XY}.

Darlington argued that the distinction between subjective decisions and mathematics should be made explicit in fairness models:

> If a conflict arises between the two goals of maximizing a test's validity and minimizing the test's discrimination against certain cultural groups, then a subjective, policy-level decision must be made concerning the relative importance of the two goals (p. 71).

According to Darlington, decision-makers should be asked to consider different goals such as low r_{CX} and high r_{XY}, and choose a value of k, indicating the value of selection of members from some subpopulation (C). Instead of predicting Y, tests would be constructed to predict (Y - kC). This model is equivalent to the regression model when k equals zero. Operationally, Darlington would add points to the scores of one group and apply the same cutting score rather than set different cutting scores for the two groups.

Einhorn and Bass (1971) Equal Risk Model

Einhorn and Bass (1971) argued that "one must take into account differences between subgroups with respect to test-criterion correlations, criterion means and variances, and differences in standard errors of estimate if one is to avoid unfair discrimination" (p. 261). That is, a test is used fairly if the risk or probability for success is equal in both groups. Their procedure involves specifying an acceptable degree of risk (constant across subgroups) and computing separate cutting scores on the predictor for each subgroup. The predictor cut-off for each group is at the maximum probability of a selection error as defined by the acceptable false positive rate (risk), given the predictor score.

According to Petersen and Novick (1976), the equal risk model is internally consistent. They found that the equal risk model is a linear function of its converse.

Measurement Invariance/Structural Equity Model

The measurement invariance/structural equity model defines equity in terms of relationships among latent factors. Meredith and Millsap (1992) defined measurement invariance to mean that the same latent variables are measured with the same degree of accuracy in each subpopulation. Gregory (1991) added the principle of structural equity to Meredith and Millsap's definition. According to Gregory,

> An unbiased or equitable test is one which, when used to select individuals with the ability to perform a task, yields equal probabilities of selection for all individuals with equal levels of ability relevant to the task, regardless of race, sex, age, etc. (p. 2).

In effect, this is the Cleary Regression Model when the latent variables, rather than the observed predictor and criterion scores are used on the X and Y axes.

The model uses structural equations modeling such as LISREL methodology to examine the factorial comparability of both the criterion and predictor domains across subgroups. Thus, it can only be tested if multiple predictor and multiple criterion data are available. The data that are demanded limit the degree to which the model can be used in research or to guide predictor development. However, it does illustrate a number of fairness issues at their most basic level.

Adverse Impact, Fairness, and Affirmative Action

As noted above, under certain conditions, the different fairness models lead to somewhat different prescriptions about how to make selection decisions within subgroups; but in a number of instances the prescriptions are neither explicit or perfectly clear. Comparisons are easiest to make when the Cleary Model holds. That is, the regression slopes, intercepts, and standard errors of measurement are the same for the subgroups. However, even if the slopes, intercepts, and errors of estimate are the same for the subgroups, there still may be a significant mean difference on the predictor. In this instance an organization may also adopt affirmative action goal and seek a strategy that trades off some decrement in aggregate predicted performance for a higher selection rate for particular subgroups.

Selection Goals

If an organization sets goals for the number or proportion of people to be selected from each group then there are a variety of ways to make the decisions, depending on the way decision-making is evaluated and the value, or utility, that is ascribed to various decision outcomes (Petersen & Novick, 1976). If maximizing aggregate predicted performance across all selectees is the goal but the organization also wishes to satisfy specific affirmative action goals, then Hunter et al. (1977) have demonstrated that top down selection within groups yields the best trade-off. However, top down selection within groups may be organizationally or politically complicated. One compromise that has been suggested is the procedure of "score banding" (Cascio, Outtz, Zedeck, & Goldstein, 1991).

Score Banding

Cascio et al. (1991) proposed banding as an alternative to top-down selection. Score banding is not a fairness model; it is a way of using test scores. The procedure involves using the standard error of measurement (SEM) and standard error of the difference between scores (SED) to form a band of scores within which scores are not statistically significantly different from each other at some alpha level. Two types of bands, sliding and fixed, have been proposed. The top, or highest, score is the referent for forming the band, and the sliding band adjusts down as top scorers are selected. The authors suggest that selection of individuals within the band can be made at random, can be based on other job-relevant criteria, or can be made to meet affirmative action goals.

Schmidt (1991) criticized the method as being inconsistent with the traditional selection goal of maximizing predicted performance for those selected. "Banding" will result in lower aggregate predicted performance than using top down selection within groups and the difference will be greater the wider the bands. Zedeck, Outtz, Cascio, and Goldstein (1991) acknowledge this be default; and, without saying so diiectly make the argument that top down selection within groups requires specific subgroup goals (i.e. quotas). Quotas also come into play when selection within bands is based on affirmative action goals. Random selection within bands does not raise the specter of quotas, can serve the general goal of affirmative action (depending upon the magnitude of the differences in scores and the width of the band), and can keep the decrease in mean predicted performance at some acceptable level (depending upon the width of the band).

Summary

It should now be evident that different models and formulae make different value-laden assumptions about fairness. Organizations that select the optimal Cleary model choose to maximize predicted job performance. The Thorndike and Cole models are suboptimal, but they are responsive to social concerns. All of the models, except the structural equity model, assume that subgroup differences on the criterion are real (i.e., do not reflect bias). It is questionable whether many of the models are "legal" under the CRA of 1991 which clearly disallows different cutting scores and score adjustments.

A Current Event: The Case of the GATB

As mentioned before, the National Academy of Sciences (NAS) chose to deviate from the Cleary model in making recommendations for the use of General Aptitude Battery Test (GATB) scores (Hartigan & Wigdor, 1989). The committee reviewed and reanalyzed GATB data to form its conclusions.

The commission found some evidence of differential validity when validity estimates for blacks were compared to those for whites. The correlation between the GATB composite (for a selected job family) and the criterion measure (supervisor's ratings) was larger for the sample of white employees than for the sample of black employees in 48 out of 72 validity studies. The average validity estimate (weighted for sample size) was .19 for white employees and .12 for black employees. On the average there were 87 blacks in each study compared to 166 whites, and thus the validity estimates based on data for blacks showed greater variability due to sampling error than did those for whites. Also, the differnce in validity could be due to differential range restriction for the two groups.

The committee examined standard errors of prediction, slopes, and intercepts to investigate differential prediction. The standard error of prediction, indicating the precision of prediction, was larger for blacks tnan for whites in 40 of the 72 studies and larger for whites than blacks in the remaining 32 studies. The slopes of the regression of the criterion scores on the GATB composite scores were significantly different in only 2 of the 72 studies ($p<.05$). Further examination of the slopes showed that there was a tendency for the slope to be greater for whites than for blacks. The committee then tested for intercept differences using data from the 70 studies without significant slope differences. Intercepts were significantly different ($p<.05$) in 26 of the 70 studies, and in every case but one intercepts were greater for white than for black employees. Finally, the committee compared predicted criterion scores based on the total-group equation with those based on subgroup equations. Both the white equation and the total-group equation tended to overpredict black criterion scores.

The committee concluded that "given the low correlation and the substantial difference in mean scores of blacks and whites on the GATB, use of the test for selection of black applicants without taking the applicant's race into account would yield very modest gains in average criterion scores but would have substantial adverse impact" (p. 185). The committee, therefore, recommended use of score adjustments that give approximately equal chances of referral to able minority applicants and able majority applicants.

The committee compared the effects of within-group percentile scoring and performance-based scoring (using the Cole, 1973, definition) in different scenarios to determine what type of adjustments would be reasonable. The score adjustment for computation of performance-based scores was to add $(1 - r^2)m$ to each minority score, where r is the correlation between test score and job performance, and m is the difference between majority- and minority-group means. The committee found that the within-group percentile scoring and performance-based scoring yielded essentially the same impact on adverse impact and resulted in equivalent drops in validity, as long as validities were modest as they are for the GATB. Consequently, the committee concluded that either method would yield essentially the same result.

After the CRA of 1991 passed, the Department of Labor (DOL) held a press conference on the status of the GATB. DOL announced that it would continue research to improve the GATB. In the interim, DOL advised states to use the GATB according to their own discretion. Therefore, it is possible that fewer organizations are now using the GATB given that DOL is not supporting its use.

Are virtually all of the suggestions for fair test use made by Thorndike, Cole, and others illegal according to the CRA of 1991? That will depend on how the CRA is interpreted, particularly on how "adjust" is defined and the extent to which predicted performance is considered. Imagine that a test overpredicts job performance of blacks. Is it unfair to use the raw score for top-down selection because by using the full group regression line one indirectly makes an upward adjustment to the predicted performance scores of blacks? Or is fairness defined only in terms of the predictor score such that any adjustment of that score, regardless of the magnitude of mean differences on the criterion or the degree of predictor-criterion correlation, would be unfair? Such questions will be decided in legal settings.

Fairness Research Issues

Classification *System* Fairness

Fairness in the military setting is more complex than selection fairness alone. Unlike the civilian sector, where job applicants are typically candidates for only one specific job, military job applicants are often candidates for more than one job. The military allocates people across jobs. During wartime, charges of unfairness are likely to arise if minorities appear to be disproportionately represented in combat jobs (e.g., Walters, 1991). Similarly, disproportionate numbers of women in administrative and clerical jobs, compared to technical jobs, can appear unfair. Also, some jobs have better advancement opportunities or civilian sector counterparts; underrepresentation of minorities in these jobs is another fairness matter. Disproportionate representation is a form of adverse impact--an outcome. It could be (and has been) controlled through minority fill rates or quotas for jobs.

For classification systems like those used by the Services it might be useful to more formally identify and elaborate all the points in the decision system at which disparate treatment could occur (e.g., high school recruitment, applicant screening, classification screening - who qualifies for what jobs or training assignments, retention rates). A flow diagram could map out some primary issues. Does the military inform youth about the kinds of educational experiences that will lead to a preferred job? Does the military seek out well-qualified minority youth? How fair is the initial selection screen? What variables enter the classification decision? What are the fairness implications of each step in a Service's classification algorithm? It is possible that there is something akin to the Cleary model that can be constructed at each stage. Such an explicit analysis would enable the Services to pinpoint problem areas and identify ways to promote opportunity while maintaining readiness.

With regard to enlisted first-tour classification, there are four major decision points in the enlisted first-tour classification model for most Services: (1) recruitment, (2) enlistment screening, (3) initial classification based on input to a person-job-match (PJM) algorithm prior to enlistment, and (4) final classification via a second PJM system. As an example, of the systems perspective, consider one portion of the larger decision system, the pre-enlistment PJM algorithm. And, for illustrative purposes consider the Air Force's pre-enlistment PJM algorithm, Procurement Management Information System (PROMIS). PROMIS generates a relative payoff index that reflects the value of assigning the recruit to each job. Five components enter the PROMIS payoff algorithm to form the payoff index (with a maximum of 1,000 points):

(1) variable fill versus aptitude/difficulty, 600 points,
(2) predicted technical school success, 50 points,
(3) occupational area preference (for Mechanical, Administrative, General, or Electrical occupations), 180 points,
(4) minority/non-minority, 70 points, and
(5) constant fill, 100 points (Pina, 1988).

Variable fill is an index of the Air Force's needs at a particular point in time (i.e., number of personnel needed and the time remaining to fill the AFS). The aptitude/difficulty subcomponent matches individual aptitude to the level of aptitude required by the job (i.e., job difficulty). Variable fill and Aptitude/Difficulty interact such that aptitude/difficulty receives a larger allocation of the 600 points, if the Air Force's need for recruits is being met and vice versa. The technical school success component is based on regression equations for predicting technical school grades from AFQT, M, A, G, and E composites, and binary variables representing high school courses taken. The area preference component assigns points to M, A, G, and E areas in proportion to the applicant's preference. When PROMIS was originally developed the minority/ nonminority component was designed to help meet the Air Force minority representation goals set for each AFS. Our most recent information is that the minority fill component still exists in the algorithm, but receives no points (L.T. Looper, personal communication, 14 April 1992). "Constant fill" is simply a constant of 100 points added to every AFS for which the applicant is eligible.

The aptitude/difficulty, training success, and occupational preference components each lend themselves to examination via the Cleary model because they are buttressed by prediction equations. Constant fill could possibly be couched in the Cleary framework; however, eligibility is a function of several factors other than scores on the ASVAB (e.g., height, weight, strength, color-vision). Once analyzed, the individual PROMIS components would need to be considered within the larger system because a recruit's ultimate assignment to a job is an outcome of recruitment, selection, the pre-enlistment PJM system, and another PJM system used during Basic Military Training.

Validity Generalization

Another fairness research issue concerns the generalizability of results across jobs and/or organizations. Fairness models generally require information about predictor-criterion relationships and relatively large samples of minorities to be included in criterion-related validation studies. Many public and private sector organizations rely on content-validation or have insufficient numbers of minorities available to examine fairness psychometrically. It would be highly useful to develop a method of meta-analyzing prediction systems for categories of tests and jobs (e.g., obtaining population estimates of slopes and intercepts).[2] Such an analysis would enable us to better understand the conditions under which differential prediction is likely to occur.

Military Policy Issues of Fairness

Debates about combat exclusion laws/policies and use of full least squares regression weights are examples of areas where the definition of fairness is currently in question.

[2] This idea comes from a suggestion by Malcolm J. Ree.

Combat Exclusion Laws/Policies

Whether women serve in combat positions is an important question for the military selection and classification research community. If, within the next several years, women gain entree to combat jobs, the Services' will need to evaluate whether the ASVAB is an adequate classification tool for combat jobs, given the expansion of the applicant population to women.

The Presidential Commission on the Assignment of Women in the Armed Forces (1992) considered issues surrounding the selection and assignment of women in the military. The commission was highly polarized and neither advocated nor denounced a combat role for women. The Roper Organization had conducted two surveys for the Commission, one to investigate the attitude of the civilian population toward women serving in combat roles and a counterpart survey of the military population (Roper Organization Inc., August 1992, September 1992). Public opinion tended to favor allowing women to serve in direct combat jobs; indeed there was a good deal of support for requiring women to take combat positions (as opposed to making the decision voluntary). In general, the military sample opposed assignment of women to combat duty.

Use of Full Least Squares Regression Weights

In the spring of 1992, we interviewed military selection and classification research experts to identify classification research objectives and concerns (Russell, Knapp, & Campbell, 1992). One fairness issue that arose dealt with using full least squares (FLS) regression-weighted ASVAB test scores for classification. FLS weights can be negative. If applicants are told to do the best on the ASVAB, is use of negative weights fair? Some people we interviewed felt that negative weights would be unfair in a selection context, but could be used fairly in the classification setting, particularly if ASVAB instructions were modified to accurately reflect the way test scores would be used (i.e., to match individuals to jobs in accordance with abilities). Others felt that negative weighting would be unfair, regardless. This issue will take on more importance as selection and classification researchers consider FLS models more seriously.

Summary

Obviously, fairness is a value-laden concept. Different models and formulae make different assumptions about what is fair. The Cleary, or regression model, is optimal in that it maximizes predicted job performance. The Thorndike and Cole models are suboptimal, but they are responsive to social concerns. Even so, it is questionable whether any of the models are "legal" under the CRA of 1991 which clearly disallows different cutting scores and score adjustments. Clearly, no decision about the appropriate model of fairness for the Services can be made in a vacuum. Discussion among military selection and classification experts, policy experts, and civilians will continue to be necessary to reach a consistent definition of fairness for the Services.

REFERENCES

Abbe, C. N. (1968). *Statistical properties of allocation averages* (Research Memorandum 68-13). Washington, D.C.: U.S. Army Behavioral Science Research Laboratory.

American Institutes for Research (1992). *A guide to test validation*. Washington, D.C.: The ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W. H. Freeman.

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36,* 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education, 508-600.

Armor, D. J., Fernandez, R. L., Bers, K., & Schwarzbach, D. (1982). *Recruit aptitudes and Army job performance. Setting enlistment standards for infantrymen* (R-2874-MRAL). Santa Monica, CA: Rand Corporation.

Asher, H. M. (1983). *Causal modeling*. Newbury Park, CA: Sage.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Bartlett, C.J., Bobko, P. & Pine, S.M. (1977). Single-group validity: Fallacy of the facts? *Journal of Applied Psychology, 62,* 155-157.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program (Manual for version 2.0)*. Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards or criterion-referenced tests. *Review of Educational Research, 56,* 137-172.

Bloxom, B. (1992, March). Armed Services Vocational Aptitude Battery and its Use in Selection and Classification. A briefing presented to the ASVAB Review Workshop sponsored by the Defense Manpower Data Center, Washington, DC.

Bobko, P., & Bartlett, C.J. (1978). Subgroup validities: Differential definitions and differential prediction. Journal of Applied Psychology, 63, 12-14.

Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. Psychometrika, 31(4), 507-534.

Boehm, V.R. (1972). Negro-white differences in validity of employment and training selection procedures. Journal of Applied Psychology, 56, 33-39.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. Psychometrika, 51, 375-377.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. Journal of Applied Psychology, 76(6), 863-872.

Braun, H. I. (1989). Empirical Bayes methods: A tool for exploratory analysis. In R. D. Bock (Ed.) Multilevel analysis of educational data. San Diego, CA: Academic Press Inc.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65-76.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 173-195.

Brogden, H. E (1955). Least squares estimates and optimal classification. Psychometrika, 20, 249-252.

Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. Educational and Psychological Measurement, 19, 181-190.

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. Educational and Psychological Measurement, 10, 159-186.

Browne, M. W., & Cudeck, R. (in press). Alternative ways of assessing model fit. Chapter to appear in K. A. Bollen & J. S. Long (Eds.), Testing structural equation models. Beverly Hills, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. Psychological Bulletin, 101, 147-158.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. Journal of Applied Psychology, 65, 543-558.

Callender, J. C., & Osburn, H. G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. Journal of Applied Psychology, 66, 274-281.

Camara, W. J., & Laurence, J. H. (1987). Military classification of high aptitude recruits. (FR-PRD-87-21). Alexandria, VA: Human Resources Research Organization.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81-105.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette and L. M. Hough (Eds.) Handbook of industrial and organizational psychology: Volume 1 (2nd Edition). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P. (Ed.). (1986). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year (Report 813101). Alexandria, VA: U. S. Army Research Institute.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., vol. 1, pp. 687-732). Palo Alto: Consulting Psychologists Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1992). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), New developments in selection and placement. San Francisco: Jossey-Bass.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313-333.

Cascio, W.F. (1982). Applied Psychology in Personnel Management (second edition). Reston VA: Reston Publishing Company, Inc.

Cascio, W. F. (1992). Applied Psychology in Personnel Management (third edition). Englewood Cliffs, NJ: Prentice-Hall.

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. Human Performance, 4, 233-264.

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cole, N.S. (1973). Bias in selection. Journal of Educational Measurement, 10, 237-255.

Crafts, J. L., Szenas, P. L., Chia, W. J., & Pulakos, E. D. (1988). A review of models and procedures for synthetic validation for entry-level Army jobs (ARI Research Note 88-107). Alexandria, VA: U.S. Army Research Institute for the Social and Behavioral Sciences.

Cronbach, L. J., & Gleser, G. C. (1965). Psychological tests and personnel decisions (2nd ed.). Urbana, IL: U. of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302.

Darlington, R.B. (1971). Another look at "culture fairness." Journal of Educational Measurement, 8, 71-82.

Dickinson, T. L., & Wijting, J. P. (1976). Policy-capturing as a procedure for synthetic validation. Paper presented at the meeting of the Rocky Mountain Psychological Association.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. Annual Review of Psychology, 41, 417-440.

Department of Defense. (1987). Joint-Service efforts to link military enlistment standards to job performance. (Sixth Annual Report to the House Committee of Appropriations). Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel).

Dunbar, S.B. & Novick, M.R. (1988). On predicting success in training for men and women: Examples from Marine Corps clerical specialties. Journal of Applied Psychology, 73, 545-550.

Dunnette, M. D. (1982). Critical concepts in the assessment of human capabilities. In M. D. Dunnette and E. A. Fleishman (Eds.) Human performance and productivity: Human capability assessment (pp. 1-13). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Ebel, R. L. (1972). Essentials of educational measurement (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Einhorn, H.J. & Bass, A.R. (1971). Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 75, 261-269.

Fernandez, R. L., & Garfinkle, J. B. (1985). Setting enlistment standards and matching recruits to jobs using job performance criteria. Santa Monica, CA: Rand Corporation.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity [Monograph]. Journal of Applied Psychology, 72, 493-511.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Green, B. F., & Wigdor, A. K. (Eds.) (1988). Measuring job competency. Washington, DC: National Academy Press.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. Journal of Consumer Research, 5, 103-123.

Gregory, K. L. (1992). A reconsideration of bias in employment testing from the perspective of factorial invariance. Doctoral dissertation, University of California at Berkeley.

Guion, R. M. (1976). Recruiting, selecting, and job placement. In M.D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology (pp. 777-828). Chicago: Rand McNally College Publishing Company.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette and L. M. Hough (Eds.) Handbook of industrial and organizational psychology: Volume 1 (2nd Edition). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. Personnel Psychology, 18, 135-164.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazersfeld (ed.), Mathematical Thinking in the Social Sciences. New York: Columbia University Press.

Harris, D. A., McCloy, R. A., Dempsey, J. R., Roth, C. Sackett, P. R., Hedges, L. V., Smith, D. A., & Hogan, P. F. (1991). Determining the relationship between recruit characteristics and job performance: A methodology and a model. (FR-PRD-90-17). Alexandria, VA: Human Resources Research Organization.

Hartigan, J.A. & Wigdor, A.K. (Eds.) (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, D.C.: National Academy Press.

Hatch, R. S., Pierce, M. B., & Fisher, A. H. (1968). Development of a computer-assisted recruit assignment system (COMPASS II). Rockville, MD: Decision Systems.

Hedges, L. V. (1988). The meta-analysis of test validity studies: Some new approaches. In H. Wainer and H. I. Braun (Eds.) Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Helms, J.E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? American Psychologist, 47, 1083-1101.

Hogan, P. F., & Smith, D. A. (1991). Entry standards for military service: A cost/performance tradeoff model. Proceedings of the 33rd Annual Conference of the Military Testing Association (pp. 132-138). San Antonio, TX.

Hollenbeck, J. P., & Whitemer, E. M. (1988). Criterion-related validation for small sample contexts: An integrated approach to synthetic validity. Journal of Applied Psychology, 73, 536-544.

Horst, P. (1954). A technique for the development of a differental prediction battery. Psychological Monographs: General and Applied, 68, 9, 1-31.

Houston, W.M. & Novick, M.R. (1987). Race-based differential prediction in Air Force technical training programs. Journal of Educational Measurement, 24, 309-320.

Hull, C. L. (1928). Aptitude testing. Yonkers, H.Y.: World Book.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance Measurement and Theory (pp. 257-266). Hillsdale, NJ: ErLbaum.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340-362.

Hunter, J. E. & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychological Bulletin, 83, 1053-1071.

Hunter, J. E. & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In E.A. Fleishman and M.D. Dunnette (Eds.), Human performance and productivity, Vol. 1: Human capability assessment. Hillsdale, NJ: Erlbaum.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research finding across studies. Beverly Hills, CA: Sage.

Hunter, J.E., Schmidt, F.L., & Hunter, R.F. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. Psychological Bulletin, 86, 721-735.

Hunter, J.E., Schmidt, F.L., & Rauchenberger, J.M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. Journal of Applied Psychology, 62, 245-260.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R. M., & Keller-McNulty, S. (1991). Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests. In A. K. Wigdor and B. F. Green, Jr. (Eds.), Performance assessment for the workplace (Volume 2, pp. 258-304) . Washington, DC: National Academy Press.

Jaeger, R. M., & Busch, J. C. (1984). A validation and standard-setting study of the general knowledge and communication skills tests of the National Teacher Examinations. Final report. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina.

James, L. R. (1973). Criterion models and construct validity for criteria. Psychological Bulletin, 80(1), 75-83.

James, L. R., Demaree, R. G., & Mulaik, S. A. (1986). A note on validity generalization procedures. Journal of Applied Psychology, 70, 56-65.

Jensen, A.R. (1980). Bias in mental testing. New York: Free Press.

Johnson, R. M. (1974). Trade-off analysis of consumer values. Journal of Marketing Research, 11, 121-127.

Johnson, C.D., & Zeidner, J. (1990). Classification utility: Measuring and improving benefits in matching personnel to jobs (IDA Paper P-2240). Alexandria, VA: Institute for Defense Analysis.

Johnson, C.D., & Zeidner, J. (1991). The economic benefits of predicting job performance: Vol. II classification efficiency. New York: Praeger.

Johnson, C.D., & Zeidner, J., & Leaman, J.A. (1991). Improving classification efficiency by restructuring Army job families (TR 947). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, C.D., & Zeidner, J., & Scholarios, D. (1990). Improving the classification efficiency of the Armed Services Vocational Aptitude Battery Through the use of Alternative test selection indices. (IDA Paper P-2427). Alexandria, VA: Institute for Defense Analysis.

Jöreskog, K. G. (1966). Testing a simple structure hypothesis in factor analysis. Psychometrika, 31, 165-178.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. Psychometrika, 32, 443-482.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology, 23, 121-145.

Jöreskog, K. C., & Sörbom, D. (1981). LISREL VI user's guide: Analysis of linear structural relationships by the method of maximum likelihood (4th ed.). Uppsala, Sweden: University of Uppsala.

Jöreskog, K. G., & Sörbom, D. (1989). LISREL VII user's reference guide (First edition). Mooresville, IN: Scientific Software, Inc.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative-aptitude-treatment interaction approach to skill acquisition. Journal of Applied Psychology, 74, 657-690.

Knapp, D.K., & Campbell, J.P. (1992). Building a joint-service classification research roadmap: Criterion-related issues. Alexandria, VA: Human Resources Research Organization.

Knapp, D.K., Russell, T.R., & Campbell, J.P. (1992). Building a joint-service classification research roadmap: Job analysis methodologies. Alexandria, VA: Human Resources Research Organization.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Konieczny, F. B., Brown, G. N., Hutton, J., & Stewart, J. E. (1990). Enlisted personnel allocation system: Final report. (ARI Technical Report 902). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Science.

Kriewall, T. E. (1972). Aspects and applications of criterion-referenced tests. Downers Grove, IL: Institute for Educational Research, April, 1972. (ERIC No. ED 063 333, 27 pp).

Kroeker, L. P. (1989). Personnel classification/assignment models. In M. F. Wiskoff & G. M. Rampton (Eds.) Military personnel measurement: Testing, assignment, evaluation. New York, NY: Praeger.

Kroeker, L. P. (1988). Extending the Navy classification model. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.) Linking military enlistment standards to job performance: Report of a workshop. Committee on the Performance of Military Personnel. National Research Council. Washington, DC: National Academy Press.

Kroeker, L. P., & Folchi, J. (1984). Classification and assignment within PRIDE (CLASP) system: Development and evaluation of an attrition component. (NPRDC TR 84-40). San Diego: Navy Personnel Research and Development Center.

Kroeker, L. P., & Rafacz, B. A. (1983). CLASP: A recruit assignment model. (NPRDC TR 84-9). San Diego: Navy Personnel Research and Development Center.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! Intelligence, 14, 389-433.

Lawshe, C. H. (1952). Employee selection. Personnel Psychology, 5, 31-34.

Leaman, J. A. (1992, August). Restructuring job families to improve potential classification efficiency. Paper presented at the 100th Annual Convention of the American Psychological Association, Washington, DC.

Leighton, D. L., Kageff, L. L., Mosher, G. P., Gribben, M. A., Faneuff, R. S., Demetriades, E. T., & Skinner, M. J. (1992). Measurement of productive capacity: A methodology for Air Force enlisted specialties (AL-TP-1992-0029). Brooks Air Force Base, TX: Armstrong Laboratory.

Linn, R.L. (1973). Fair test use in selection. Review of Educational Research, 43, 139-161.

Linn, R.L. (1975). Test bias and the prediction of grades in law school. Journal of Legal Education, 27, 297-323.

Linn, R.L. (1978). Single group validity, differential validity, and differential prediction. Journal of Applied Psychology, 3, 507-512.

Linn, R.L. (in press). Fair test use: Research and policy. Proceedings of the Army Research Institute Selection and Classification Conference. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1983). A comparative study of standard-setting methods (Research Report 83-38). Princeton, NJ: Educational Testing Service.

Martin, C.J. (1992, March). Computerized testing. A briefing presented to the ASVAB Review Workshops sponsored by the Defense Manpower Data Center. Washington, DC.

Mayberry, P. W. (1987). Developing a competency scale for hands-on measures of job proficiency (CRC 570). Alexandria, VA: Center for Naval Analyses.

McCloy, R. A. (1990). A New Model of Job Performance: An Integration of Measurement, Prediction, and Theory. Unpublished Doctoral Dissertation, University of Minnesota.

McCloy, R. A., Harris, D. A., Barnes, J. D., Hogan, P. F., Smith, D. A., Clifton, D., & Sola, M. (1992). Accession quality, job performance, and cost: A cost-performance tradeoff model (FR-PRD-92-11). Alexandria, VA: Human Resources Research Organization.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 56, 347-367.

Meredith, W. & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. Psychometrika, 57, 289-311.

Mossholder, K. W., & Arvey, R. D. (1984). Synthetic validity: A conceptual and Comparative Review. Journal of Applied Psychology, 69, 322-333.

More normal nonsense. (1989, July). Fortune, p. 118.

Naylor, J. C. (1983). Modeling performance. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), Performance measurement and theory (pp. 299-305). Hillsdale, NJ: Erlbaum.

Nedelsky, L. (1954). Absolute grading sstandards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J. J., Lipner, R. S., & Langdon, L. O. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Nord, R. D., & Kearl, C. E. (1990). Estimating cost-effective recruiting missions: A profit maximizing approach. Alexandria, VA: U. S. Army Research Institute.

Nord, R. D., & Schmitz, E. J. (1989). Estimating performance and utility effects of alternative selection and classification policies. In J. Zeidner and C.D. Johnson (Eds.) (1989) The economic benefits of predicting job performance (IDA Paper P-2241). Alexandria, VA: Institute for Defense Analyses.

Osburn, H. G., Callender, J. C., Greener, J. M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. Journal of Applied Psychology, 68, 115-122.

Peterson, N. G., & Bownas, D. A. (1982). Skill, task structure, and performance acquisition. In M.D. Dunnette and E.A. Fleishman (Eds.) Human performance and productivity: Human capability assessment (pp. 49-105). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Peterson, N. G., Owens-Kurtz, & Rosse (1991). Formation of job performance prediction equations and evaluation of their validity. In L. L. Wise, N. G. Peterson, R. G. Hoffman, J. P. Campbell, & J. M. Arabian (Eds.) (1991). Army Synthetic Validity Project: Report of phase III results. ARI- TR-922. Alexandria, VA: U.S. Army Research Institute for the Social and Behavioral Sciences.

Peterson, N. G., Wise, L. L., & Campbell, J. P. (1991). Summary and discussion. In L. L. Wise, N. G. Peterson, R. G. Hoffman, J. P. Campbell, & J. M. Arabian (Eds.) (1991). Army Synthetic Validity Project: Report of phase III results. ARI- TR-922. Alexandria, VA: U.S. Army Research Institute for the Social and Behavioral Sciences.

Pina, M. (1988). Air Force person-job match: Non-prior service enlisted classification. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.) Linking military enlistment standards to job performance: Report of a workshop. Committee on the Performance of Military Personnel. National Research Council. Washington, DC: National Academy Press.

Pina, M. (1974). The assignment of airmen by solving the transportation problem (AFHRL-TR-74-58). Lackland AFB, TX: Human Resources Laboratory.

Pina, M., Jr., Emerson, M. S., Leighton, D. L., & Cummings, W. (1988). Processing and Classification of Enlistees (PACE) system payoff algorithm development (AFHRL-TP-87-41). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Poggio, J. P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document 249 267).

Presidential Commission on the Assignment of Women in the Armed Forces (1992). Report to the President. Washington, DC: U.S. Government Printing Office.

Primoff, E. S. (1955). Test selection by job analysis: The J-Coefficient, what it is, how it works (Test Technical Series, No. 20). Washington, DC: U. S. Civil Service Commission.

Pulakos, E., Wise, L., Arabian, J., Heon, S., & Delaplane, S. K. (1989). A review of procedures for setting job performance standards. Washington, DC: American Institutes for Research.

Roach, B. W. (1984). Decision-Theoretic approach to personnel selection: A review (AFHRL-TP-84-19; AD-A143 388). Brooks AFB, TX: Air Force Human Resources Laboratory.

Roper Organization Inc. (1992, August). Attitudes regarding the assignment of women in the armed forces: The public perspective. Washington, D.C.: Author.

Roper Organization Inc. (1992, September). Attitudes regarding the assignment of women in the armed forces: The military perspective. Washington, D.C.: Author.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosse, R.L. & Peterson, N.G. (1991). An investigation of the use of least squares validity estimators and correction formulas when population values are available for predictor intercorrelations. Addendum to Volume I of the Army Synthetic Validity Project: Report of Phase III Results (Report 922). Alexandria, VA: U. S. Army Research Institute.

Russell, T.L., Knapp, D.K., & Campbell, J.P. (1992). Building a joint-service classification research roadmap: Defining research objectives (HumRRO IR-PRD-92-10). Alexandria, VA: Human Resources Research Organization.

Russell, T.L., Reynolds, & Campbell, J.P (Eds.) (1992). Building a joint-service classification research roadmap: Individual differences measurement. Alexandria, VA: Human Resources Research Organization.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73, 482-486.

Sadacca, R., Campbell, J. P., White, L. A., & Difazio, A. S. (1989). Weighting criterion components to develop composite measures of job performance (Report 838). Alexandria, VA: U. S. Army Research Institute.

Schmidt, F.L. (1991). Why all banding procedures in personnel selection are logically flawed. Human Performance, 4, 265-277.

Schmidt, F. L. (1988). Validity Generalization and the future of criterion-related validity. In H. Wainer and H. I. Braun (Eds.) Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L., Hunter, J. E., Croll, P. R., & McKenzie, R. C. (1983). Estimation of employment test validities by expert judgment. Journal of Applied Psychology, 68, 590-601.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. Journal of Applied Psychology, 71, 432-439.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. P. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. Personnel Psychology, 37, 407-422.

Schmitz, E. J. (1988). Improving personnel performance through assignment policy. In B. F. Green, H. Wing, & A. K. Wigdor (Eds.) Linking military enlistment standards to job performance: Report of a workshop. Committee on the Performance of Military Personnel. National Research Council. Washington, DC: National Academy Press.

Scholarios, D. (1992, August). A comparison of predictor selection methods for maximizing potential classification efficiency. Paper presented at the 100th Annual Convention of the American Psychological Association, Washington, DC.

Sigmon, G. L., & Halpin, G. (1984, April). Application of judgmental standard setting procedures to vocational evaluation competency statements by rehabilitation field personnel and educators. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Skakun, E. N., & Kling, S. (1980). Comparability of methods of standards setting. Journal of Educational Measurement, 17, 229-235.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the Validation and Use of Personnel Selection Procedures. (Third Edition) College Park, MD: Author.

Sorenson, R. C. (1965). Optimal allocation of enlisted men - Full regression equations versus aptitude area scores (Technical Research Note 163; AD 625 224). Washington, D.C.: U.S. Army Personnel Research Office.

Sorenson, R. C. (1967). Amount of assignment information and expected performance of military personnel (TR 1152; AD 649 907). Washington, DC: U.S. Army Personnel Research Office.

Statman, M. A. (1992, August). Developing optimal predictor equations for differential job assignment and vocational counseling. Paper presented at the 100th Annual Convention of the American Psychological Association, Washington, DC.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

Steiger, J. H., & Lind, J. (1980). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tett, B. P., Jackson, D. N., & Rothstein, M. R. (1991). Personality measures as predictors of job performance: A meta-analytic review. Personnel Psychology, 44, 703-742.

Thomas, H. (1990). A likelihood-based model for validity generalization. Journal of Applied Psychology, 75, 13-20.

Thorndike, R.L. (1971). Concepts of culture fairness. Journal of Educational Measurement, 8, 63-70.

Torgerson, W. S. (1958). Theory and methods of scaling. New York: John Wiley & Sons.

Trattner, M. H. (1982). Synthetic validity and its application to the uniform guidelines validation requirements. Personnel Psychology, 35, 383-397.

von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. American Psychologist, 47, 244-253.

Walters, R. (1991, March). African-American participation in the All Volunteer Force: Lessons from the Persian Gulf Crisis. Testimony before the U.S. House of Representatives Committee on Armed Services.

Ward, J. H., Jr. (1977, August). Creating mathematical models of judgment processes: From policy-capturing to policy-specifying (AFHRL-TR-77-47, AD-AO48 983). Brooks AFB, TX: occupation and Manpower Research Division, Air Force Human Resources Laboratory.

Weitz, J. (1961). Criteria for criteria. American Psychologist, 16, 228-231.

Whetzel, D. L. (1992, August). Multidimensional screening: Comparison of a single-stage personnel selection/classification process with alternative strategies. Paper presented at the 100th Annual Convention of the American Psychological Association, Washington, DC.

Wilbourn, J. M., Valentine, L. D., & Ree, M. J. (1984). Relationship of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 to Air Force technical school final grade. (AFHRL-TP-84-8, AD-A144 213). Brooks AFB, TX: Air Force Human Resources Laboratory.

Wing, H., Peterson, N. G., & Hoffman, R. G. (1985). Expert judgments of predictor-criterion validity relationships. In Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M., Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (Report 660). Alexandria, VA: U. S. Army Research Institute, pp. 219-270.

Wise. L. L. (In Press). Goals of the selection and classification decision. Proceedings of the Army Research Institute's Selection and Classification Conference. Alexandria, VA: U.S. Army Research Institute.

Wise, L. L., Peterson, N. G., Hoffman, R. G., Campbell, J. P., & Arabian, J. M. (1991). Army synthetic validity project report of phase III results, volume I (Report 922). Alexandria, VA: U. S. Army Research Institute.

Wright, S. (1934). The method of path coefficients. Annals of Mathematical Statistics, 5, 161-215.

Zedeck, S. Outtz, J., Cascio, W.F., & Goldstein, I.L. (1991). Why do "testing experts" have such limited vision? Human Performance, 4, 297-308.

Zeidner, J. (1987, April). The validity of selection and classification procedures for predicting job performance (IDA Paper P-1977). Alexandria, VA: Institute for Defense Analyses.

Zeidner, J., & Johnson, C.D. (1989). The utility of selection for military and civilian jobs (IDA P-2239). Alexandria, VA: Institute for Defense Analyses.

Zeidner, J., & Johnson, C.D. (In Press). Is personnel classification a concept whose time has passed? Proceedings of the Army Research Institute's Selection and Classification Conference. Alexandria, VA: U.S. Army Research Institute.